

QUARTERLY OF APPLIED MATHEMATICS

Vol. XVI

OCTOBER, 1958

No. 3

ON SOME ITERATIVE METHODS FOR SOLVING ELLIPTIC DIFFERENCE EQUATIONS*

BY

HERBERT B. KELLER

Institute of Mathematical Sciences, New York University

1. Introduction. The numerical solution of boundary value problems for partial differential equations usually requires the solution of large systems of linear equations. The order, n , of such systems is essentially equal to the number of mesh points in the domain under consideration. Since direct inversion procedures require the order of n^3 operations they are not practicable, even using high speed digital computers, for reasonable meshes in two or more dimensions. Thus iterative methods for solving linear systems are of great interest as they usually require the order of n^2 operations. In addition the coefficient matrix of the system which results from the finite difference approximations has many strategically placed zeroes. However, no special account of these zeroes is taken in most direct inversions or in general iterative procedures. It is reasonable to expect that particular methods, designed in accordance with the general structure of the coefficient matrix, could further reduce the number of operations. Many such special iteration schemes have been devised and conditions on the coefficient matrix which are sufficient to insure the convergence of some of these methods have been obtained [1, 7, 9]. However there is no general comparison procedure to determine which of many possible methods is "best" in a given case.

In the present paper we formulate a family of iterative schemes for a particular class of coefficient matrices (in which the zeroes are placed as in the usual five-point Laplace difference equations). This family is defined by a generalization of the usual notion of extrapolation or over-relaxation. It is then possible to formulate the problem of finding the "best" scheme and, more important, some general theorems on the eigenvalues of these schemes are proved.

The theorems are used to define three subclasses, called complete image classes, of the general family. These classes contain many of the schemes in current use as well as generalizations of them. Thus it is shown that a variety of independently proposed and seemingly unrelated iterative methods are special cases of a general class of methods. These complete image classes are such that each of the eigenvalues of any scheme in the class is a given function of one of the eigenvalues of a particular reference scheme of the class. Thus a knowledge of the eigenvalues of the reference scheme permits, in principle, the determination of the best scheme of the given class.

A special class of equations is considered for which all the eigenvalues of each reference scheme can be explicitly written in terms of the eigenvalues of two matrices. It is then

*Received June 10, 1957. The research reported in this paper was performed under Contract AT (30-1)-1480 with the U. S. Atomic Energy Commission.

possible to determine which of the reference schemes is best and hence it is also possible to determine that scheme which is best of all those in the complete image classes.

As an example of the application of this general theory the Laplace difference equations are considered. The well-known results on rate of convergence for the Richardson, Liebmann and extrapolated Liebmann methods are immediate consequences. Analogous results are obtained for the less well-known "line" methods* which are shown to be superior.

It is clear that the methods of the present paper can be applied to more general iterative schemes than those considered. In addition many of the present results can be easily extended to problems in higher dimensions and with more general boundaries.

2. Formulation. A large class of two dimensional linear elliptic difference equations are of the form:

$$\phi_{ij} - l_{ii}\phi_{i-1,j} - r_{ii}\phi_{i+1,j} - b_{ii}\phi_{i,j-1} - t_{ii}\phi_{i,j+1} = s_{ij}; \quad (2.0)$$

within a coordinate rectangle specified by $1 < i < p$, $1 < j < q$. On the boundaries of the domain equations of the form (2.0) hold with the coefficients:

$$l_{1j} = r_{pj} = 0, \quad 1 \leq j \leq q; \quad b_{i1} = t_{iq} = 0, \quad 1 \leq i \leq p. \quad (2.1)$$

Such equations are obtained from second order† elliptic partial differential equations by applying the usual second order difference approximations on some coordinate mesh, (ξ_i, η_j) . The coefficient matrix of the resulting system, (2.0) and (2.1) must be non-singular and, with a little care in differencing [1], can be made positive definite (and symmetric if the equation is self adjoint). However, we shall assume, unless otherwise stated, only the non-singularity.

For convenience of notation and discussion we introduce, for each j , the p -dimensional column vectors:

$$\Phi_j \equiv \begin{bmatrix} \phi_{1j} \\ \phi_{2j} \\ \vdots \\ \phi_{pj} \end{bmatrix}, \quad S_j \equiv \begin{bmatrix} s_{1j} \\ s_{2j} \\ \vdots \\ s_{pj} \end{bmatrix}; \quad (2.2)$$

and the $p \times p$ order matrices:

$$L_j \equiv \begin{bmatrix} 0 & & & 0 \\ l_{2j} & 0 & & \\ l_{3j} & 0 & & \\ \cdot & 0 & & \\ \cdot & 0 & & \\ \cdot & 0 & & \\ 0 & & l_{pj} & 0 \end{bmatrix}, \quad R_j \equiv \begin{bmatrix} 0 & r_{1j} & & & 0 \\ 0 & r_{2j} & & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ 0 & & & & 0 \end{bmatrix}, \quad 1 \leq j \leq q; \quad (2.3)$$

*The origin of iterative line methods is obscure. They have been in use by Russian mathematicians for a number of years. About 1945 J. von Neuman and L. H. Thomas independently proposed line methods for parabolic difference equations. An independent investigation was initiated by M. E. Rose and the present author in 1953 and some of the results of Sec. 8 were then obtained. Peaceman and Ratchford have studied their applicability to parabolic difference equations and double-sweep iterative solutions of the Laplace difference equations.

†The five point scheme implied by (2.0) assumes no mixed partial derivatives in the equations.

$$B_i \equiv \begin{pmatrix} b_{1i} & 0 \\ & b_{2i} \\ & \ddots \\ 0 & b_{pi} \end{pmatrix}, \quad 1 < j \leq q; \quad T_i \equiv \begin{pmatrix} t_{1i} & 0 \\ & t_{2i} \\ & \ddots \\ 0 & t_{pi} \end{pmatrix}, \quad 1 \leq j < q.$$

The system (2.0-1) can now be written as

$$\begin{aligned} (I - L_1 - R_1)\Phi_1 - T_1\Phi_2 &= S_1; \\ -B_1\Phi_{i-1} + (I - L_i - R_i)\Phi_i - T_i\Phi_{i+1} &= S_i, \quad 2 \leq i \leq q-1; \\ -B_q\Phi_{q-1} + (I - L_q - R_q)\Phi_q &= S_q. \end{aligned} \quad (2.4)$$

Here we have introduced the identity matrix, I , which is always assumed to be of the same order as the square matrices to which it is added.

Further simplification is obtained by introducing the $(p \times q)$ -dimensional column vectors* (or q -dimensional compound vectors):

$$\Phi \equiv \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_q \end{bmatrix} \equiv \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \vdots \\ \phi_{pq} \end{bmatrix}, \quad S \equiv \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_q \end{bmatrix} \equiv \begin{bmatrix} s_{11} \\ s_{21} \\ \vdots \\ s_{pq} \end{bmatrix}; \quad (2.5)$$

and the $[(pq) \times (pq)]$ -order matrices (or $[q \times q]$ -order compound matrices):

$$\begin{aligned} L &\equiv \begin{pmatrix} L_1 & 0 \\ L_2 & \ddots \\ 0 & L_q \end{pmatrix}, & R &\equiv \begin{pmatrix} R_1 & 0 \\ R_2 & \ddots \\ 0 & R_q \end{pmatrix}, \\ B &\equiv \begin{pmatrix} 0 & 0 \\ B_2 & 0 \\ B_3 & 0 \\ \ddots & \ddots \\ 0 & B_q & 0 \end{pmatrix}, & T &\equiv \begin{pmatrix} 0 & T_1 & 0 \\ 0 & T_2 & \ddots \\ \ddots & \ddots & \ddots \\ 0 & 0 & T_q \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \quad (2.6)$$

The system of linear equations (2.0-1), or (2.4) now becomes

$$M\Phi \equiv (I - L - R - B - T)\Phi = S. \quad (2.7)$$

Similar formulations may be introduced for more general boundaries and in higher dimensions. In particular, if the boundaries are composed of coordinate segments, the

*The vector Φ of (2.5) determines an "ordering" [1] of the unknowns, Φ_{ij} ; but this order need bear no relationship to the sequence in which the iterative computations are carried out.

matrices L_i and R_i remain square but of different orders while the matrices B_i and T_i become rectangular. Another pair of "neighbors", ϕ_{iik+1} , appear in (2.0) with each unit increase of the dimension and, correspondingly, additional pairs of matrices must be introduced in a manner similar to those of (2.6).

3. General single sweep iterations. We summarize here some terminology and known results for a class of iterative methods for solving (2.7); this class is defined as follows: Let the coefficient matrix, M , be written as

$$M = N - P, \quad (3.0)$$

where $|N| \neq 0$. We call this a "splitting" of the coefficient matrix and the system (2.7) becomes

$$N\Phi = P\Phi + S, \quad (3.1)$$

with the formal solution

$$\Phi = (N - P)^{-1}S = (I - N^{-1}P)^{-1}N^{-1}S. \quad (3.2)$$

The iterative procedure is defined, starting from some arbitrary guess, $\Phi^{(0)}$, at the solution vector, by the recursion

$$N\Phi^{(\nu)} = P\Phi^{(\nu-1)} + S. \quad (3.3)$$

Thus in one sweep through the mesh a new iterate is obtained. Applying (3.3) recursively yields for the ν th iterate

$$\Phi^{(\nu)} = [I + (N^{-1}P) + (N^{-1}P)^2 + \cdots + (N^{-1}P)^{\nu-1}]N^{-1}S + (N^{-1}P)^{\nu}\Phi^{(0)}. \quad (3.4)$$

Thus [2] $\Phi^{(\nu)} \rightarrow \Phi$ as $\nu \rightarrow \infty$, for arbitrary $\Phi^{(0)}$, if and only if

$$\lim_{\nu \rightarrow \infty} (N^{-1}P)^{\nu} = 0. \quad (3.5)$$

This condition is satisfied provided some appropriate norm [3] of $(N^{-1}P)$, say the spectral norm, is < 1 .

This result is more frequently obtained by introducing the sequence of error vectors

$$E^{(\nu)} \equiv \Phi - \Phi^{(\nu)}, \quad (3.6)$$

which, by (3.1) and (3.3), must satisfy the homogeneous recursion

$$NE^{(\nu)} = PE^{(\nu-1)}, \quad \nu \geq 1. \quad (3.7)$$

The eigenvalues, λ_k , of $N^{-1}P$ are the (pq) roots of the characteristic equation

$$|\lambda N - P| = 0. \quad (3.8)$$

If they are distinct* there then exists [4] a complete set of eigenvectors, e_k , satisfying

$$\lambda_k Ne_k = Pe_k, \quad (3.9)$$

which span the (pq) -dimensional vector space. Thus any initial error, $E^{(0)}$, has a unique expansion in these eigenvectors of the form

$$E^{(0)} = \sum_{k=1}^{pq} a_k e_k. \quad (3.10)$$

*It is sufficient here to assume that all the elementary divisors [4] of $(N^{-1}P)$ are simple.

By (3.7) and (3.9) the above yields, for the ν th error vector,

$$E^{(\nu)} = \sum_{k=1}^{pq} \lambda_k' a_k e_k. \quad (3.11)$$

In order that $\Phi^{(\nu)} \rightarrow \Phi$ it is necessary and sufficient that $E^{(\nu)} \rightarrow 0$. Thus by (3.11) and the completeness of the eigenvectors, convergence is equivalent, for arbitrary initial error*, to

$$\lambda_{\max} \equiv \max_k |\lambda_k| < 1. \quad (3.12)$$

Let us require that the most slowly decaying component in the ν th error, (3.11), be reduced by at least 10^{-m} , where $m > 0$. Then we must have $\lambda_{\max}^{\nu} \leq 10^{-m}$ and the number of iterations required is bounded by

$$\nu \geq -m/\log \lambda_{\max} \equiv m/R. \quad (3.13)$$

This result is valid only when (3.12) is satisfied and then $R \equiv -(\log \lambda_{\max})^{-1}$ is called the rate of convergence. This quantity is useful in comparing different iterative methods as the number of iterations required for some specified convergence criterion varies as R^{-1} .

If the elementary divisors of $N^{-1}P$ are not simple, condition (3.12) still suffices for convergence but the bound (3.13) does not apply. To obtain such a bound we assume the elementary divisor of largest order corresponding to $\lambda_{\max} = |\lambda_1|$ say, is of order $r + 1$. Then the expansion (3.10) must be replaced by [4]

$$E^{(\nu)} = \sum_{k=1}^{r+1} \left[\sum_{s=0}^{r+1-k} \binom{\nu}{s} \lambda_1^{\nu-s} a_{s+k} \right] e_k + \sum_{k=r+2}^{pq} \lambda_k' a_k e_k \quad (3.14)$$

provided $\nu \geq (r + 1)$ (and assuming all other divisors to be simple). The largest decay factor is now given by, for $\nu \geq (r + 1)\lambda_{\max} + r$,

$$\lambda_{\max}^{\nu-r} \binom{\nu}{r};$$

and to reduce the error by at least 10^{-m} requires

$$\lambda_{\max}^{\nu-r} \binom{\nu}{r} \leq 10^{-m}.$$

An approximate bound on the number of iterations which may be obtained from this inequality, is

$$\nu \geq m'/R + r + (r/2R) \log [(m'/R + r)m'/R], \quad (3.15)$$

where $m' \equiv m - \log r!$. The number of iterations required for convergence is now not simply proportional to R^{-1} but this result is asymptotically (for $m \rightarrow \infty$) equivalent to (3.13). For practical computations the discrepancy may be significant.

A large rate of convergence should not be the only criterion in evaluating iterative methods. Rather a measure of the time required (by human and/or machine effort) to achieve the desired accuracy should be employed. This time is essentially proportional

*For a special initial error in which the amplitude of some particular component vanishes, say $a_1 = 0$, the corresponding eigenvalue, λ_1 , is unrestricted. However, even if such initial distributions could be determined, computations with roundoff would undoubtedly introduce this component at an early stage of the iterations and it could become arbitrarily large if (3.12) is violated.

to the total number of arithmetic operations (properly weighted). Thus if we let N_{op} be the operational count required for each iteration [i.e. to solve (3.3)], the total time is proportional to [assuming (3.13) to be valid],

$$T \equiv N_{op}/R. \quad (3.16)$$

The "best" of the single sweep methods is that for which T is minimized. Of course in more general procedures a similar criterion should be employed.

4. Special single sweep iterations and generalized extrapolation. With the above notions in mind we consider the special class of splittings

$$N(\gamma) \equiv \gamma_0 I - \gamma_1 L - \gamma_2 R - \gamma_3 B - \gamma_4 T, \quad P(\gamma) \equiv N(\gamma) - M, \quad (4.0)$$

where the real numbers γ_i are restricted by the conditions that

$$\text{a)} \quad |N(\gamma)| \neq 0, \quad \text{b)} \quad \gamma_1 \gamma_2 \gamma_3 \gamma_4 = 0. \quad (4.1)$$

This defines a subset Γ of the five dimensional Euclidian space of all points γ . The iterations are defined by

$$N(\gamma)\Phi^{(r)} = P(\gamma)\Phi^{(r-1)} + S. \quad (4.2)$$

Thus any point $\gamma \in \Gamma$ determines an iterative scheme given by (4.0) and (4.2); we shall sometimes refer to this as the scheme γ . The class of schemes, Γ , is important since it includes many known and frequently used schemes while the data arrangement required for all of these schemes is well suited for automatic computing machines. Furthermore, the solution of the system (4.2) is obtained explicitly when $\gamma \in \Gamma$ in one sweep over the mesh by solving either two-term or three-term recursions. In particular if either $\gamma_0 \gamma_1 \gamma_2 \neq 0$ or $\gamma_0 \gamma_1 \gamma_4 \neq 0$ three-term recursions are introduced along horizontal or vertical mesh lines. The solution of these recursions may be reduced, by the well-known factorization of a Jacobi matrix, to the evaluation of two, two-term recursions along the appropriate lines. If two or more of the $\gamma_i \neq 0$ the method of sweeping the mesh to solve (4.2) in one sweep is partially determined (i.e. the sweep must start at a particular corner or with some line next to a boundary).

The operational counts, $N_{op}(\gamma)$, required to solve (4.2) for any of the schemes $\gamma \in \Gamma$ vary at most by a factor less than two. The minimum number of operations needed is four multiplications and five additions at each mesh point and the maximum required is seven multiplications and six additions (neglecting the operations done only once in factoring the matrix of Jacobi form). The special subclasses of Γ introduced in Sec. 6 require at most six multiplications and five additions at each point. Thus in the remainder of the paper we shall assume the operational counts to be almost equal and in seeking the "best" scheme we shall consider only the eigenvalues.

The eigenvalues of a scheme such as (4.0) to (4.2) are the roots λ of the characteristic equation

$$|\lambda N(\gamma) - P(\gamma)| = 0;$$

or explicitly, of the equation

$$\Delta \equiv |g_0 I - g_1 L - g_2 R - g_3 B - g_4 T| = 0, \quad (4.3)$$

where

$$g_i \equiv \gamma_i(\lambda - 1) + 1, \quad 0 \leq i \leq 4.$$

Each root, $\lambda_k(\gamma)$, is thus a function of the scheme, γ , and a problem naturally suggested is to find a $\gamma^* \in \Gamma$ such that

$$\lambda_{\max}(\gamma^*) = \min_{\gamma \in \Gamma} (\max_k |\lambda_k(\gamma)|). \quad (4.4)$$

The solution of this problem would yield, essentially, the best iterative method of the class considered. The usual notion of extrapolation (or over-relaxation) of the scheme γ^* is meaningless, by virtue of (4.4), since no improvement on rate of convergence can be obtained. In fact the usual extrapolation techniques may be viewed as attempts to find the solution of (4.4) when γ is further restricted to lie on some particular curve in Γ . This is in fact shown to be the case in Sec. 6 and the curve, in the usual extrapolation procedures, is a straight line. The problem posed in (4.4) is thus a generalization of extrapolation in which the values of four extrapolation parameters are to be obtained [since, by (4.1b), Γ consists of four four-dimensional subspaces].

In order to obtain some idea of the possible behavior of $\lambda(\gamma)$ we consider schemes γ on the "diagonal" line

$$\gamma_0 = \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 1/\alpha. \quad (4.5)$$

These points do not lie in Γ and, obviously, taking $\alpha = 1$ implies direct inversion of M . However, to gain insight, we use (4.5) in (4.3) and obtain the characteristic equation:

$$\left| \left[\frac{1}{\alpha} (\lambda - 1) + 1 \right] M \right| = 0.$$

Since, by assumption, $|M| \neq 0$, there is only one eigenvalue (with multiplicity pq) and it is

$$\lambda = 1 - \alpha.$$

Thus along the diagonal line (4.5) all schemes converge for $0 < \alpha < 2$ (i.e. $\gamma_i > 1/2$) and diverge otherwise (i.e. $\gamma_i < 1/2$). Only one iteration is required for $\alpha = 1$, as then $\lambda = 0$. It might be suspected that the best scheme in Γ is obtained by taking a point nearest this unit point [$\alpha = 1$ in (4.5)]. That this is not the case is a simple consequence of the results of Sec. 6. The eigenvalues of the above example become arbitrarily large in absolute value as $\alpha \rightarrow \pm\infty$ and the point γ approaches the origin; the roots approach unity as $\alpha \rightarrow \pm 0$ and the point γ recedes to infinity. It should also be observed that these eigenvalues are independent of the matrix M .

5. Some properties of the eigenvalues, $\lambda(\gamma)$. We present here some theorems which can be used to compare the eigenvalues of various schemes γ . All of these results are simple consequences of the

Fundamental theorem: Let L , R , B and T be arbitrary matrices of the form (2.6), (2.3). Then for any non-zero scalars x and y ,

$$|M| \equiv |I - L - R - B - T| \equiv \left| I - xL - \frac{1}{x}R - yB - \frac{1}{y}T \right|. \quad (5.0)$$

Proof. Let the elements of M be M_{rs} , where $1 \leq r, s \leq N = pq$. Then each term in the formal expansion of $|M|$ is given by [5]

$$\pm M_{1,\pi(1)} M_{2,\pi(2)} \cdots M_{r,\pi(r)} \cdots M_{N,\pi(N)}, \quad (5.1)$$

where π is one of the $N!$ permutations of the first N integers. Let each point (i, j) of

the original rectangular mesh (see Fig. 1) be identified with a unique integer $r = (j-1)p + i$, and represent any permutation π by the N vectors from r to $\pi(r)$ on this mesh. By the definitions (2.6), (2.3) and (5.0), $M_{r, \pi(r)} \neq 0$ only if $\pi(r) = r$ or the point corresponding to $\pi(r)$ is an adjacent neighbor of the point r . Thus the only permutations which lead to non-vanishing terms (5.1) are those whose geometric representation is composed entirely of unit vectors in the $(\pm i)$ - and $(\pm j)$ -directions and null vectors. However, any permutation is a product of disjoint cycles [5] and the representation of a cycle is a closed path on the mesh. Thus for non-vanishing cycles there are the same number of unit vectors in the $(+i)$ -direction as in the $(-i)$ -direction and similarly for the $(\pm j)$ -directions (see Fig. 1). Since $M_{r, \pi(r)}$ is an element of L if $\pi(r) = r - 1$, and an element of R if $\pi(r) = r + 1$, there are as many factors from L as from R in each non-vanishing cycle. A similar result is true of factors from B and T . Thus each non-vanishing term in the expansion of the right hand determinant in (5.0) is independent of x and y and the proof is complete.

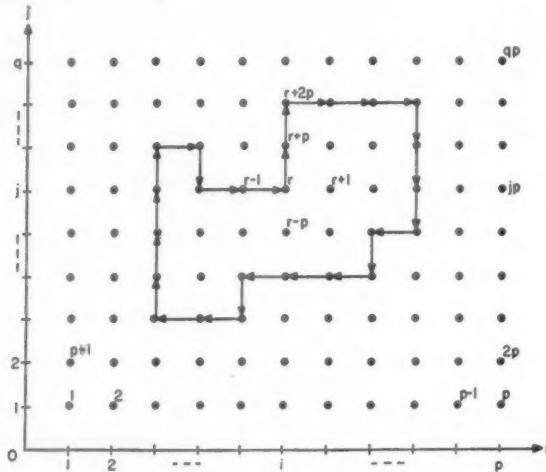


FIG. 1. Geometric representation of a non-vanishing cycle. In the permutation of $n \rightarrow \pi(n)$, which this cycle is a factor the cycle is given by:

$$\pi(r) = r + p, \quad \pi(r + p) = r + 2p, \dots, \quad \pi(r - 1) = r.$$

The geometric form of the above proof was suggested by D. Ludwig. Another proof has been given by B. Friedman [8]. However, the geometric proof clearly indicates that the same result can be proved for a matrix which represents the difference equations on any connected region bounded by coordinate segments. In higher dimensions the analogous theorem, obtained by adding two more matrices to M for each additional dimension, is easily proved. A special case of the above theorem, namely that obtained by setting $x = y$ in (5.0) has been proven by D. Young [1]. Many of the remaining results in this paper hold for any matrix M which can be written in the form (2.7) such that (5.0) is satisfied for arbitrary x and y .

The first, almost obvious, consequence of the fundamental theorem is contained in *Theorem I*. Let $\lambda(\gamma) = \lambda(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ be any eigenvalue of the scheme γ . Then for all γ :

$$\lambda(\gamma) = \lambda(\gamma_0, \gamma_2, \gamma_1, \gamma_3, \gamma_4) = \lambda(\gamma_0, \gamma_1, \gamma_2, \gamma_4, \gamma_3). \quad (5.2)$$

Proof. The eigenvalues $\lambda(\gamma)$ are the roots of the characteristic equation (4.3). However, by the fundamental theorem we have

$$\Delta = \left| g_0 I - xg_1 L - \frac{1}{x} g_2 R - yg_3 B - \frac{1}{y} g_4 T \right|,$$

and taking $x = (g_2/g_1)^{1/2}$, $y = (g_4/g_3)^{1/2}$ the above yields*

$$\Delta = | g_0 I - (g_1 g_2)^{1/2} (L + R) - (g_3 g_4)^{1/2} (B + T) |. \quad (5.3)$$

Thus the roots of $\Delta = 0$ are invariant under the interchanges $\gamma_1 \leftrightarrow \gamma_2$ and $\gamma_3 \leftrightarrow \gamma_4$.

In terms of iterative procedures of the class (4.2), Theorem I indicates that various ways of sweeping the mesh yield identical rates of convergence. In terms of the subspace Γ the theorem states that there are certain two-dimensional planes with respect to which the eigenvalues are symmetric. Thus the volume of Γ which must be searched for a best scheme, γ^* , is reduced by a factor of 1/4. Furthermore it is of interest to note that along the two-dimensional planes $\gamma_1 = \gamma_2$ and $\gamma_3 = \gamma_4$ the eigenvalues $\lambda(\gamma)$ must have relative maxima or minima with respect to the appropriate pair of coordinates.

A somewhat more general result which contains Theorem I as a special case is contained in

Theorem II. Let $\lambda'(\gamma')$ be some particular eigenvalue of the scheme γ' . For any scheme γ let $\lambda(\gamma)$ be a function such that

$$\left(\frac{g_0}{g'_0} \right)^2 = \frac{g_1 g_2}{g'_1 g'_2} = \frac{g_3 g_4}{g'_3 g'_4} \equiv \xi^2 \neq 0, \quad (5.4)$$

where g'_i and g_i are respectively the functions of (λ', γ') and (λ, γ) defined in (4.3). Then $\lambda(\gamma)$ is an eigenvalue of the scheme γ .

Proof. Form the determinant Δ of (4.3). Then as in the proof of Theorem I we obtain (5.3). Using (5.4) this becomes

$$\Delta = \xi^{pq} | g'_0 I - (g'_1 g'_2)^{1/2} (L + R) - (g'_3 g'_4)^{1/2} (B + T) | = \xi^{pq} \Delta',$$

by another application of the fundamental theorem to the scheme γ' . However, since $\lambda'(\gamma')$ is an eigenvalue of γ' we have $\Delta' = 0$. Thus $\Delta = 0$ and $\lambda(\gamma)$ must be an eigenvalue of γ .

The schemes γ and γ' , and the eigenvalues $\lambda(\gamma)$ and $\lambda'(\gamma')$ of the theorem will be called images of each other. We sometimes refer to γ' or $\lambda'(\gamma')$ as the reference scheme or reference eigenvalue respectively. It is clear from (5.4) that the relationship of being images is a transitive one. Some consequences of Theorem II are examined in the remaining sections.

Another theorem which is quite useful for some special difference equations is

Theorem III. Let the matrices (2.6) be such that $(L + R)(B + T) = (B + T)(L + R)$. Then as is well known these matrices have common eigenvectors; let ρ_k and μ_k be the eigenvalues of $(L + R)$ and $(B + T)$ respectively, corresponding to the common eigenvector e_k . Then each eigenvalue of any scheme γ is a root, for some k , of the (at most 4th degree) equation**

$$g_0 - (g_1 g_2)^{1/2} \rho_k - (g_3 g_4)^{1/2} \mu_k = 0. \quad (5.5)$$

*The chosen branches of the square roots are arbitrary and hence all future applications of (5.3) could be written in any of four ways.

**As pointed out in the proof of Theorem I there are four such equations which could be used. This point is clarified in Sec. 7 where more of the consequences of the hypothesis are examined.

Proof. Let $\lambda(\gamma)$ be some fixed but as yet unspecified eigenvalue of the scheme γ . Then from (4.3), $\Delta = 0$, and as in the proof of Theorem I, we have from (5.3)

$$|g_0I - (g_1g_2)^{1/2}(L + R) - (g_3g_4)^{1/2}(B + T)| = 0. \quad (5.6)$$

Thus the matrix in (5.6) is singular and has a zero eigenvalue. Applying this matrix to e_k we see that

$$\xi_k \equiv g_0 - (g_1g_2)^{1/2}\rho_k - (g_3g_4)^{1/2}\mu_k$$

is an eigenvalue belonging to the eigenvector e_k . Using all the e_k we obtain all of the eigenvalues of the matrix in (5.6). However at least one $\xi_k = 0$. Since $\lambda(\gamma)$ was any eigenvalue of the scheme γ the theorem follows.

This theorem is applied generally in Sec. 7 and to the usual Laplace difference equations in Sec. 8.

6. The complete image classes. The pairs of image schemes defined by Theorem II are such that only one eigenvalue of any γ need be the image of one eigenvalue of the reference scheme, γ' . However, of special interest are those schemes each of whose eigenvalues is the image of a corresponding eigenvalue of some particular reference scheme. Such classes of schemes are called complete image classes and we proceed to obtain three of them. The intersections of these classes with the subspace Γ is grouped into five subspaces: Γ_A , Γ_B , $\Gamma_{B'}$, Γ_C and $\Gamma_{C'}$, of which $\Gamma_{B'}$ and $\Gamma_{C'}$ are considered uninteresting.

The class Γ_A . In order that any two schemes γ and γ' be images, (5.4) must be satisfied by the corresponding pair of image eigenvalues. However, let us seek first schemes such that

$$\frac{g_1g_2}{g'_1g'_2} \equiv \frac{g_3g_4}{g'_3g'_4} \quad (6.0)$$

is an identity in the indeterminates λ and λ' . This requires

$$\text{a) } \begin{cases} \gamma_1\gamma_2 = \gamma_3\gamma_4 \\ \gamma_1 + \gamma_2 = \gamma_3 + \gamma_4 \end{cases}, \quad \text{b) } \begin{cases} \gamma'_1\gamma'_2 = \gamma'_3\gamma'_4 \\ \gamma'_1 + \gamma'_2 = \gamma'_3 + \gamma'_4 \end{cases}. \quad (6.1)$$

Now let λ' be any eigenvalue of a scheme γ' which satisfies (6.1b). Then if γ is any scheme satisfying (6.1a) every root λ of

$$\left(\frac{g_0}{g'_0}\right)^2 = \frac{g_1g_2}{g'_1g'_2} \quad (6.2)$$

is an eigenvalue of γ by Theorem II. Since (6.0) is satisfied for any λ and λ' , each equation (6.2) obtained for a different eigenvalue λ' of γ' determines at least one eigenvalue λ of the scheme γ . Thus all schemes γ satisfying (6.1a) belong to the same complete image class, say class A. Since the conditions (6.1a) and (6.1b) are identical any $\gamma \in A$ may be used as the particular reference scheme. We call the chosen reference scheme γ_A and take it to be

$$\gamma_A : \gamma_0 = 1, \quad \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.$$

This is a simultaneous displacement method commonly known as Richardson's method. Thus each of the eigenvalues of any class A scheme is the image of one of the Richardson eigenvalues.

The class A schemes of interest are those contained in Γ . From (4.1) and (6.1) we obtain the set of all such schemes, Γ_A , and they are listed in Table A. The real parameters α and β are to be restricted such that (4.1a) is satisfied. The set Γ_A lies on four two-dimensional planes in Γ , one of which is, say: $\gamma_2 = \gamma_4 = 0, \gamma_1 = \gamma_3$. By the symmetry

	γ_0	γ_1	γ_2	γ_3	γ_4
Γ_A	$1/\alpha$	0	$1/\beta$	$1/\beta$	0
	$1/\alpha$	$1/\beta$	0	$1/\beta$	0
	$1/\alpha$	0	$1/\beta$	0	$1/\beta$
	$1/\alpha$	$1/\beta$	0	0	$1/\beta$
γ_A	1	0	0	0	0

TABLE A

property of the eigenvalues, expressed in Theorem I, we need examine only one of these four planes. However using any $\gamma \in \Gamma_A$ and γ_A in (6.2) yields

$$[\lambda - (1 - \alpha)]^2 = \frac{\alpha^2}{\beta} \lambda_A^2 [\lambda - (1 - \beta)]. \quad (6.3)$$

This equation furnishes the mapping of the eigenvalues λ_A of γ_A (Richardson) onto the eigenvalues λ of any of the schemes in Γ_A .

If we choose $\alpha = \beta = 1$ in any $\gamma \in \Gamma_A$ the resulting scheme is a successive displacement method commonly known as a Liebmann scheme, in the present connection, or more generally as the Gauss-Seidel method (see Table I). The mapping (6.3) yields

$$\lambda = \lambda_A^2, \quad \text{or} \quad \lambda = 0. \quad (6.4)$$

Thus some λ_A go into zero and others into their squares. These schemes converge, as is well known [7], if the Richardson scheme converges, and the rate of convergence is twice as large (see Sec. 3).

If we set $\beta = 1$ in any $\gamma \in \Gamma_A$ the resulting scheme is a successive overrelaxation [1] method where α is the overrelaxation parameter (see Table I); this method is frequently called the extrapolated Liebmann scheme [6]. The schemes γ , in this case, lie along four lines in Γ_A , for example: $\gamma_2 = \gamma_4 = 0, \gamma_1 = \gamma_3 = 1$. The mapping (6.3) becomes

$$[\lambda - 1 + \alpha]^2 = \alpha^2 \lambda_A^2 \lambda, \quad (6.5)$$

which has been studied thoroughly [1].

Taking $\alpha = \beta$ in any $\gamma \in \Gamma_A$ yields a different successive overrelaxation method; this corresponds to using successive displacements (Liebmann) over the entire mesh and then extrapolating (or interpolating) the provisional new iterate with the old iterate. The mapping (6.3) becomes, in this case,

$$\lambda = 1 - \alpha(1 - \lambda_A^2), \quad (6.6)$$

TABLE I
Some standard methods as complete images.

Name, Type and Class	$\gamma_0 \gamma_1 \gamma_2 \gamma_3 \gamma_4$	g_0	g_1	g_2	$g_3 g_4$	$\left(\frac{g_0}{g_1}\right)^2$	$\frac{g_1 g_2}{g'_1 g'_2}$	$\frac{g_1 g_4}{g'_1 g'_4}$
Richardson (Simultaneous displacement)	$A \begin{smallmatrix} 1 & 0 & 0 & 0 & 0 \end{smallmatrix}$	ρ	1	1	1 1
Liebmann (G-S) (Successive displacement)	$A \begin{smallmatrix} 1 & 1 & 0 & 1 & 0 \end{smallmatrix}$	λ	λ	1	$\lambda \ 1$	$(\lambda/\rho)^2$	λ	λ
Extrapolated Liebmann (Successive overrelaxation)	$A \begin{smallmatrix} 1 & 1 & 0 & 1 & 0 \\ \alpha & & & & \end{smallmatrix}$	$\xi - 1 + \alpha$	ξ	1	$\xi \ 1$	$\left(\frac{\xi - 1 + \alpha}{\alpha \rho}\right)^2$	ξ	ξ
Line-Richardson (Group simultaneous displacement)	$B \begin{smallmatrix} 1 & 1 & 1 & 0 & 0 \end{smallmatrix}$	ν	ν	ν	1 1
Line-Liebmann (Group successive displacement)	$B \begin{smallmatrix} 1 & 1 & 1 & 1 & 0 \end{smallmatrix}$	ρ	ρ	ρ	$(\rho/\nu)^2$	$(\rho/\nu)^2$	ρ	ρ
Extrapolated Line-Lieb- mann (Group successive overrelaxation)	$B \begin{smallmatrix} 1 & 1 & 1 & 1 & 0 \\ \alpha & \alpha & \alpha & \alpha & \end{smallmatrix}$	$\mu - 1 + \alpha$	$\mu - 1 + \alpha$	$\mu - 1 + \alpha$	$\mu \ 1$	$\left(\frac{\mu - 1 + \alpha}{\alpha \nu}\right)^2$	$\left(\frac{\mu - 1 + \alpha}{\alpha \rho}\right)^2$	μ

which is easily analyzed. If the λ_A are all real and $\lambda_{\max, A} < 1$, then λ_{\max} is minimized by taking

$$\alpha = [1 - \frac{1}{2}(\lambda_{\max, A}^2 + \lambda_{\min, A}^2)]^{-1},$$

where $\lambda_{\min, A} \equiv \min_k |\lambda_k(\gamma_A)|$. This method is the analogue of the case treated at the end of Sec. 4; it is called a full-mesh extrapolation.

If all of the eigenvalues λ_A are real and $\lambda_{\max, A} < 1$ the general mapping (6.3) can be analyzed. It is found, in this case, that the best of the class A schemes is extrapolated Liebmann. However, in other cases, it seems possible to improve upon the Liebmann extrapolations. The analysis of this mapping has been done by K. Gordis and will be reported in a future paper.

The class Γ_B . Proceeding as in the previous case we now require

$$\left(\frac{g_0}{g'_0}\right)^2 = \frac{g_1 g_2}{g'_1 g'_2}$$

to be an identity in λ and λ' . This is equivalent to

$$\text{a)} \quad \gamma_0 = \gamma_1 = \gamma_2, \quad \text{b)} \quad \gamma'_0 = \gamma'_1 = \gamma'_2. \quad (6.7)$$

Using (4.1b) these relations yield the indicated classes Γ_B and $\Gamma_{B'}$ of Table B.

	γ_0	γ_1	γ_2	γ_3	γ_4
Γ_B	$1/\alpha$	$1/\alpha$	$1/\alpha$	$1/\beta$	0
	$1/\alpha$	$1/\alpha$	$1/\alpha$	0	$1/\beta$
$\Gamma_{B'}$	0	0	0	$1/\alpha$	$1/\beta$
γ_B	1	1	1	0	0

TABLE B

The canonical reference scheme, γ_B , included in the table is a simultaneous line-displacement method; by its analogy with γ_A we shall call it a line-Richardson method. The classes of schemes Γ_B and $\Gamma_{B'}$ lie on three two-dimensional planes in Γ . As before if we take λ_B to be any eigenvalue of γ_B then by Theorem II any root λ of

$$\frac{g_3 g_4}{g'_3 g'_4} = \left(\frac{g_0}{g'_0}\right)^2$$

is an eigenvalue of $\gamma \in \Gamma_B$ (or $\Gamma_{B'}$) of the table. Of course the appropriate $\gamma \in \Gamma_B$ (or $\Gamma_{B'}$) is to be used in the g_i . Using $\gamma \in \Gamma_B$ and γ_B of the table this yields

$$[\lambda - (1 - \alpha)]^2 = \frac{\alpha^2}{\beta} \lambda_B^2 [\lambda - (1 - \beta)]; \quad (6.8)$$

the same mapping (6.3) as in the Γ_A schemes. However, the reference schemes γ_A and γ_B are not images of each other and thus we cannot, in general compare their eigenvalues. In Secs. 7 and 8 special cases are considered in which λ_A and λ_B may be compared; it is then clear that some schemes in Γ_B are better than the best scheme in Γ_A .

All of the simplifications (6.4-6) are shown to apply for $\gamma \in \Gamma_B$ by choosing the

same special α and β as were chosen for $\gamma \in \Gamma_A$ (see Table I). Thus we may call $\gamma = (1, 1, 1, 1, 0)$ "line-Liebmann", and $\gamma = (1/\alpha, 1/\alpha, 1/\alpha, 1, 0)$ "extrapolated line-Liebmann" and their eigenvalues have exactly the same relationships to the line-Richardson eigenvalues as the eigenvalues of ordinary Liebmann and extrapolated Liebmann have to those of ordinary Richardson.

The remaining class of schemes, $\Gamma_{B'}$, yield the mapping

$$[\lambda - (1 - \alpha)][\lambda - (1 - \beta)] = \alpha\beta\lambda_B^2. \quad (6.9)$$

This result has been examined and is in general found to be inferior to that of (6.8).

The class Γ_C . This class is determined exactly as the previous one by interchanging (γ_1, γ_2) with (γ_3, γ_4) . The mappings (6.8) and (6.9) are obtained for the corresponding classes Γ_C and $\Gamma_{C'}$. The canonical reference scheme is $\gamma_C = (1, 0, 0, 1, 1)$, a line-Richardson method, which is not an image of γ_A or γ_B . In the special cases of the next two sections it is possible to compare λ_B and λ_C and thus to determine which of the classes, Γ_B or Γ_C contains the best scheme.

The essential difference between Γ_B and Γ_C schemes is the direction of the line along which three-term recursions must be solved. In all Γ_B schemes they are parallel to the direction of increasing i and in Γ_C parallel to the direction in which j increases.

7. Reference eigenvalues for a special class of equations. We consider here those systems of the form (2.2) to (2.7) for which $(L + R)$ and $(B + T)$ commute. Then the fundamental theorem and Theorem III are valid. However, before applying these results we shall examine some other very important consequences of commutativity for matrices of the indicated forms.

By using the forms (2.6) in

$$(L + R)(B + T) = (B + T)(L + R)$$

we obtain the conditions

$$\begin{aligned} (L_i + R_i)B_j &= B_j(L_{j-1} + R_{j-1}), & 2 \leq j \leq q, \\ (L_i + R_i)T_j &= T_j(L_{j+1} + R_{j+1}), & 1 \leq j \leq q-1. \end{aligned}$$

These conditions are necessary and sufficient for commutativity. With no loss in generality* we may assume $|B_i| \neq 0$ and $|T_i| \neq 0$ in which case the above conditions become

$$(L_i + R_i) = \begin{cases} B_j(L_{j-1} + R_{j-1})B_j^{-1}, & 2 \leq j \leq q, \\ T_j(L_{j+1} + R_{j+1})T_j^{-1}, & 1 \leq j \leq q-1. \end{cases} \quad (7.0)$$

Since the $(L_i + R_i)$ are all similar they have the same eigenvalues and elementary divisors; and as the order of these matrices is $p \times p$ there are at most p distinct eigenvalues ρ_i , $1 \leq i \leq p$. From the form (2.6) of $(L + R)$ we see that its eigenvalues are just the ρ_i , each of whose multiplicity has a factor q [i.e. if ρ_i has multiplicity r_i in $(L_i + R_i)$ then it has multiplicity $r_i q$ in $(L + R)$]. If the ρ_i are all distinct (or belong to simple elementary divisors) it can be shown from (7.0) that for some scalars a_i ,

$$B_{i+1} = a_i T_i^{-1}, \quad 1 \leq i \leq q-1. \quad (7.1)$$

*Only the boundary conditions may introduce singular B_i and T_i for elliptic difference equations. However, these cases may be eliminated, depending on the boundary conditions, by either not including the boundary values as unknowns (as in Sec. 8) or by requiring the difference analogue of the equations to hold at boundary points.

By changing the ordering (2.2), (2.5) from the original row-ordering to a column-ordering (which is equivalent to a similarity transformation of L , R , B and T) we find, as above, that the eigenvalues μ_i of $(B + T)$ are the eigenvalues of p similar $q \times q$ matrices. As eigenvalues of $(B + T)$ each μ_i has a multiplicity which is a multiple of p and there are at most q such eigenvalues. It is well known that commuting matrices have common eigenvectors. Thus in the above case there exist common eigenvectors e_{ii} such that

$$(L + R)e_{ii} = \rho_i e_{ii}, \quad (B + T)e_{ii} = \mu_i e_{ii}, \quad (7.2)$$

for each ρ_i and μ_i belonging to different elementary divisors of $(L + R)$ and $(B + T)$ respectively.

To determine some properties of the ρ_i and μ_i we consider the characteristic equations of which they are the roots. Exactly as in the proof of the fundamental theorem we see that

$$|\rho I - (L_i + R_i)| = \left| \rho I - xL_i - \frac{1}{x}R_i \right|$$

for arbitrary $x \neq 0$. Thus taking $x = -1$ we find that if ρ is an eigenvalue of $(L_i + R_i)$ then so is $-\rho$. Furthermore if p is odd there is at least one zero eigenvalue and additional zero eigenvalues must occur in pairs. Analogous results* are true of the eigenvalues μ of $(B + T)$.

We are now in a position to apply Theorem III. We select first $\gamma_A \equiv (1, 0, 0, 0, 0)$ and using (7.2) we may replace the pair (ρ_k, μ_k) of the theorem by the pair (ρ_i, μ_i) to obtain from (5.5)

$$\lambda_{ii,A} = \rho_i + \mu_i. \quad (7.3A)$$

Similarly using $\gamma_B \equiv (1, 1, 1, 0, 0)$ we get

$$\lambda_{ii,B} = \frac{\mu_i}{1 - \rho_i}, \quad (7.3B)$$

and, finally, with $\gamma_C \equiv (1, 0, 0, 1, 1)$ Theorem III yields

$$\lambda_{ii,C} = \frac{\rho_i}{1 - \mu_i}. \quad (7.3C)$$

We have introduced an obvious double subscript notation for the λ . These are all of the roots of the corresponding schemes since (7.3) holds for all ρ_i and μ_i . Using other schemes γ in (5.5) would yield many other eigenvalues explicitly. However, any of the class Γ_A , Γ_B or Γ_C scheme eigenvalues are obtained by using (7.3) in (6.2) etc.

As a further condition let us assume that the ρ_i and μ_i are all real. (A sufficient condition for this would be that $(L + R)$ and $(B + T)$ are symmetric. Then $B_{i+1} = T_i$ and by (7.1) the T_i and B_i must be constants times the identity. Similar results would hold for the transformed $(L + R)$ block matrices.) Then by the parity of the eigenvalues ρ_i and μ_i (i.e. since $\max_i |\rho_i| = \max_i \rho_i$) we obtain from (7.3A)

$$\lambda_{\max,A} = \rho_{\max} + \mu_{\max}. \quad (7.4A)$$

*The properties of the eigenvectors (7.2) and the sign parity of the eigenvalues explains the apparent ambiguity pointed out in Theorem III, in which any of four seemingly different equations could have been used. They are not different but correspond to different labeling of the eigenvalues.

Furthermore, if $\rho_{\max} < 1$ and $\mu_{\max} < 1$ we get from (7.3B, C):

$$\lambda_{\max, B} = \frac{\mu_{\max}}{1 - \rho_{\max}} ; \quad (7.4B)$$

$$\lambda_{\max, C} = \frac{\rho_{\max}}{1 - \mu_{\max}} . \quad (7.4C)$$

Thus we may easily compare these principal eigenvalues to obtain

$$a) \quad \rho_{\max} + \mu_{\max} \leq 1 \Rightarrow \lambda_{\max, A} \geq (\lambda_{\max, B}, \lambda_{\max, C}), \quad (7.5)$$

$$b) \quad |\rho_{\max} - 1/2| \leq |\mu_{\max} - 1/2| \Rightarrow \lambda_{\max, B} \leq \lambda_{\max, C} .$$

Since the mappings of the eigenvalues of the schemes in Γ_A , Γ_B and Γ_C from the corresponding image scheme eigenvalues are the same, the best scheme will be found in that class for which λ_{\max} is smallest. The relations (7.5) may thus be used to determine the best class. It should be noted, from (7.5a), that in the present case if the Richardson method converges ($\lambda_{\max, A} < 1$) then the line-Richardson methods converge faster, and if Richardson does not converge neither do the line-Richardson methods.

8. Laplace's equation. As an example we consider $\Delta^2\phi = 0$ in the rectangle $0 \leq x \leq 1, 0 \leq y \leq 1$; with $\phi =$ (given function) on the boundary. Using the mesh

$$x_i = i\Delta x, \quad \Delta x \equiv \frac{1}{p}, \quad 0 \leq i \leq p; \quad (8.0)$$

$$y_j = j\Delta y, \quad \Delta y \equiv \frac{1}{q}, \quad 0 \leq j \leq q;$$

and the difference approximations $\partial^2\phi/\partial x^2 \doteq 1/\Delta x^2 (\phi_{i+1,i} - 2\phi_{ii} + \phi_{i-1,i})$, etc., at (x_i, y_j) , we obtain the difference equations

$$\phi_{ii} = \theta_x[\phi_{i-1,i} + \phi_{i+1,i}] + \theta_y[\phi_{i,i-1} + \phi_{i,i+1}], \quad \begin{cases} 1 \leq i \leq p-1 \\ 1 \leq j \leq q-1 \end{cases}. \quad (8.1)$$

Here $\phi_{0,i}$, $\phi_{p,i}$, $\phi_{i,0}$ and $\phi_{i,q}$ are the given boundary values and

$$\theta_x \equiv \frac{\Delta y^2}{2(\Delta x^2 + \Delta y^2)}, \quad \theta_y \equiv \frac{\Delta x^2}{2(\Delta x^2 + \Delta y^2)}, \quad (\theta_x + \theta_y = \frac{1}{2}). \quad (8.2)$$

Equations (8.1) may be written in the notation of Sec. 2 where $l_{ii} = r_{ii} = \theta_x$, $b_{ii} = t_{ii} = \theta_y$ and the inhomogeneous terms come from the boundary values; the system is of order $[(p-1)(q-1)]^2$. The matrices $(L + R)$ and $(B + T)$ then commute and are symmetric, and as proved in Sec. 7, their eigenvalues are the same as those of respectively,

$$\theta_x \begin{bmatrix} 0 & 1 & & & 0 \\ 1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & 1 \\ 0 & & & 1 & 0 \end{bmatrix}_{(p-1) \times (p-1)} \quad \text{and} \quad \theta_y \begin{bmatrix} 0 & 1 & & & 0 \\ 1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & 1 \\ 0 & & & 1 & 0 \end{bmatrix}_{(q-1) \times (q-1)}.$$

The eigenvalues of these matrices are easily found to be

$$\begin{aligned}\rho_i &= 2\theta_x \cos(i\pi/p), \quad 1 \leq i \leq p-1; \\ \mu_j &= 2\theta_y \cos(j\pi/q), \quad 1 \leq j \leq q-1.\end{aligned}\tag{8.3}$$

The parity properties deduced in Sec. 7 are seen to be satisfied as $\rho_i = -\rho_{p-i}$ and $\mu_i = -\mu_{q-i}$. Thus the principal eigenvalues are obtained for $i = j = 1$, and we have

$$\begin{aligned}\rho_{\max} &= 2\theta_x \cos \pi/p \approx 2\theta_x - \theta_x(\pi/p)^2, \\ \mu_{\max} &= 2\theta_y \cos \pi/q \approx 2\theta_y - \theta_y(\pi/q)^2,\end{aligned}\tag{8.4}$$

where the approximate values hold for $p \gg \pi$ and $q \gg \pi$.

Let us consider first iterative solutions of (8.1) by means of the canonical reference schemes of Sec. 6 (all of which are simultaneous displacement methods). The principal eigenvalues of these schemes are obtained, since Theorem III applies, by using (8.4) in (7.3); then, as above, retaining only terms up to second order in $1/p$ and $1/q$ we have

$$\begin{aligned}\lambda_{\max,A} &\approx 1 - \pi^2 \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), \quad [\text{Richardson}] \\ \lambda_{\max,B} &\approx 1 - \frac{\pi^2}{2\theta_x} \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), \quad [\text{Horizontal line-Richardson}] \\ \lambda_{\max,C} &\approx 1 - \frac{\pi^2}{2\theta_y} \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), \quad [\text{Vertical line-Richardson}].\end{aligned}\tag{8.5}$$

In this approximation we see that the direction of sweep, for minimum λ_{\max} of the above line-methods, is determined only by the mesh ratio and does not depend upon the number of mesh points in the x or y -directions (since both p and q were assumed large). If the number of points is "small" the analogous criterion is obtained by using the exact eigenvalues (8.4) in (7.5). The rates of convergence (Sec. 3) of the above schemes are easily compared by recalling that $-\log(1-\epsilon) \approx \epsilon$ for small ϵ . If the mesh is square $\theta_x = \theta_y = 1/4$ and the line methods converge twice as fast as the ordinary Richardson (to second order in $1/p$). If the mesh is rectangular then $\theta_x < 1/4$ or $\theta_y < 1/4$ and by sweeping in the proper direction further improvement is obtained (the implicit equations should come from lines parallel to the direction of largest mesh spacing).

The successive displacement methods corresponding to the above reference schemes may be taken as (see Table I)

$$\gamma_a = (1, 1, 0, 1, 0), \quad \gamma_b = (1, 1, 1, 1, 0), \quad \gamma_c = (1, 1, 0, 1, 1).\tag{8.6}$$

The eigenvalues of these schemes are related to the corresponding reference eigenvalues (since they are complete image schemes) by (6.3) with $\alpha = \beta = 1$. We note as in Sec. 6 that $\lambda = 0$ may become an eigenvalue of the schemes (8.6) independently of the values of λ' . The remaining roots become $\lambda = (\lambda')^2$, and corresponding to (8.5) we get, up to second order in $1/p$ and $1/q$,

$$\begin{aligned}\lambda_{\max,a} &\approx 1 - 2\pi^2 \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), \quad [\text{Liebmann}] \\ \lambda_{\max,b} &\approx 1 - \frac{\pi^2}{\theta_x} \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), \quad [\text{Horizontal line-Liebmann}] \\ \lambda_{\max,c} &\approx 1 - \frac{\pi^2}{\theta_y} \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), \quad [\text{Vertical line-Liebmann}].\end{aligned}\tag{8.7}$$

Of course the rates of convergence are now rigorously twice those of the reference scheme rates. The best sweep direction is determined as in the previous case, and again the best line method converges at least twice as fast as the ordinary Liebmann (to second order in $1/p$ and $1/q$).

The successive overrelaxation schemes which are usually applied to the above are

$$\gamma_a = (1/\alpha, 1, 0, 1, 0), \quad \gamma_b = (1/\alpha, 1/\alpha, 1/\alpha, 1, 0), \quad \gamma_c = (1/\alpha, 1, 0, 1/\alpha, 1/\alpha), \quad (8.8)$$

and now (6.3) holds in each case with $\beta = 1$. As is well known [1] for γ_a above (which is extrapolated Liebmann) and hence for all these cases, λ_{\max} will be a minimum when α is chosen as the smaller root of

$$\lambda_{\max, X}^2 \alpha^2 - 4\alpha + 4 = 0, \quad (8.9A)$$

where $X = A, B$ or C .

[That is, solve (6.3) for λ and set the discriminant equal to zero when $\lambda_A = \lambda_{\max, A}$.] Then we have all $|\lambda| = \lambda_{\max}$ where

$$\lambda_{\max} = \alpha - 1. \quad (8.9B)$$

Using the reference eigenvalues (8.5) in the above we obtain for the schemes (8.8), up to second order in $1/p$ and $1/q$,

$$\begin{aligned} \lambda_{\max, a} &= 1 - 2\pi \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right)^{1/2} + 2\pi^2 \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), & [\text{Extrapolated Liebmann}] \\ \lambda_{\max, b} &= 1 - (2/\theta_x)^{1/2} \pi \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right)^{1/2} + \frac{\pi^2}{\theta_x} \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right), & [\text{Extrapolated horizontal line-Liebmann}] \\ \lambda_{\max, c} &= 1 - (2/\theta_y)^{1/2} \pi \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right)^{1/2} + \frac{\pi^2}{\theta_y} \left(\frac{\theta_x}{p^2} + \frac{\theta_y}{q^2} \right). & [\text{Extrapolated vertical line-Liebmann}]. \end{aligned} \quad (8.10)$$

There is an order of magnitude improvement in the rates of convergence of the extrapolated schemes over the previous schemes. That is, since $(\theta_x/p^2 + \theta_y/q^2) \equiv \Delta x^2 \Delta y^2 (\Delta x^2 + \Delta y^2)^{-1}$, the present rates are $O(\Delta x)$ or $O(\Delta y)$ while from (8.5) and (8.7) the rates are $\theta(\Delta x^2)$ or $\theta(\Delta y^2)$. The line schemes now improve the convergence rate by a factor $2^{1/2}$ for a square mesh or larger for rectangular meshes swept in the proper direction.

BIBLIOGRAPHY

1. D. Young, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Am. Math. Soc. **76**, 92–111 (1954)
2. L. V. Kantorovich, *Functional analysis and applied mathematics*, N.B.S. Rept. **1509** (1952), translation by C.D. Benster; Uspekhi Matematicheskikh Nauk **3**, 89–185 (Russian)
3. A. S. Householder, *On the convergence of matrix iterations*, Oak Ridge Natl. Lab., ORNL-1883 (June 1955)
4. J. H. M. Wedderburn, *Lectures on Matrices*, Am. Math. Soc. Colloq. Pubs. **17** (1934)
5. G. Birkhoff and S. MacLane, *A survey of modern algebra*, Macmillan Co., New York, 1946
6. S. Frankel, *Convergence rates of iterative treatments of partial differential equations*, MTAC **4**, 65–75 (1950)
7. H. Geiringer, *On the solution of systems of linear equations by certain iteration methods*, Reissner Anniversary Vol., Ann Arbor, Mich., 365–393, 1949
8. B. Friedmann, *The iterative solution of elliptic difference equations*, AEC Computing Facility, N.Y.U., Rept. NYO-7698 (1957)
9. L. F. Richardson, *The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stress in a masonry dam*, Phil. Trans. Roy. Soc. (London) **210A**, 307–357 (1910)

ON THE SOLUTION OF CERTAIN DIFFERENTIAL EQUATIONS
BY CHARACTERISTIC FUNCTION EXPANSIONS*

BY

TSE-SUN CHOW

Research Staff, General Motors Corporation, Warren, Michigan

1. In this article we seek for the solution of the differential equation

$$p_0 \frac{\partial^2 u}{\partial x^2} + p_1 \frac{\partial u}{\partial x} + p_2 u = \frac{\partial u}{\partial t}, \quad (1.1)$$

where p_0, p_1, p_2 are functions of x , (x real) with the initial condition

$$u(x, t) = u_0(t), \quad t = 0, \quad (1.2)$$

and the following boundary conditions

$$\alpha_1 u(a, t) + \alpha_2 u(b, t) + \alpha_3 u_x(a, t) + \alpha_4 u_x(b, t) = f(t), \quad (1.3)$$

$$\beta_1 u(a, t) + \beta_2 u(b, t) + \beta_3 u_x(a, t) + \beta_4 u_x(b, t) = g(t), \quad (1.4)$$

where α_i, β_i are constants¹, and $u_x = \partial u / \partial x$. Equation (1.1) is a differential equation of the second order of the parabolic type. Special cases of problems of this kind occur in heat conduction and diffusion, usually with simpler types of boundary conditions. Since the boundary conditions (1.3) and (1.4) are non-homogeneous and time-dependent, the method of separation of variables cannot be used directly. We shall first use a transformation² to remove the non-homogeneous boundary conditions and then separate the variables. This results in the well known Sturm-Liouville system and the solution of (1.1) ... (1.4) will be sought as expansions of the characteristic functions of this system. With arbitrary values of α_i, β_i the resulting Sturm-Liouville system is in general not self-adjoint, and the characteristic functions are not orthogonal. Yet, it is known in the theory of differential equations that if we introduce the adjoint system, the characteristic functions of the two systems will be bi-orthogonal, i.e., the characteristic function of one system for one particular characteristic number will be orthogonal to all the characteristic functions of the other system with the exception of one of the same characteristic number. In carrying out the expansion procedure to find the solution of (1.1) ... (1.4) we shall make use of this bi-orthogonality relationship and shall show in the final solution how the time-dependent functions, $f(t)$ and $g(t)$ are related to the boundary forms complementary to those of the adjoint system. These are given by (4.3), (4.4) and (4.5)³.

*Received Feb. 1, 1957; revised manuscript received June 10, 1957.

The author wishes to thank Professor R. C. F. Bartels of the University of Michigan for comments. He is also indebted to the referee for pointing out a somewhat misleading statement in the manuscript.

¹We assume that $\sigma_{12}, \sigma_{13}, \sigma_{24}, \sigma_{34} \neq 0$, where $\sigma_{12} = \alpha_1 \beta_2 - \alpha_2 \beta_1$ etc.

²A homogeneous differential equation with non-homogeneous boundary conditions is equivalent to a non-homogeneous differential equation with homogeneous boundary conditions [1].

³The theory of non-self-adjoint boundary value problems and the associated expansions of functions in terms of bi-orthogonal systems of characteristic functions does not appear nearly as well known as the theory of self-adjoint systems. Readers are referred to Coddington and Levinson [2] for information on this subject.

2. To remove the non-homogeneous boundary conditions we make the substitution

$$u(x, t) = \xi(x, t) + \sum_{i=1}^2 X_i T_i$$

in Eqs. (1.1) ··· (1.4) where X_i and T_i are functions of x and t respectively. We get

$$p_0 \frac{\partial^2 \xi}{\partial x^2} + p_1 \frac{\partial \xi}{\partial x} + p_2 \xi - \frac{\partial \xi}{\partial t} = \sum_{i=1}^2 \{X_i T'_i - T_i L(X_i)\}, \quad (2.1)$$

$$\xi(x, 0) = u_0(x) - \sum_{i=1}^2 X_i T_i(0), \quad (2.2)$$

$$\begin{aligned} & \alpha_1 \xi(a, t) + \alpha_2 \xi(b, t) + \alpha_3 \xi_x(a, t) + \alpha_4 \xi_x(b, t) \\ & + \sum_{i=1}^2 \{\alpha_1 X_i(a) T_i(t) + \alpha_2 X_i(b) T_i(t) + \alpha_3 X'_i(a) T_i(t) + \alpha_4 X'_i(b) T_i(t)\} = f(t), \end{aligned} \quad (2.3)$$

$$\begin{aligned} & \beta_1 \xi(a, t) + \beta_2 \xi(b, t) + \beta_3 \xi_x(a, t) + \beta_4 \xi_x(b, t) \\ & + \sum_{i=1}^2 \{\beta_1 X_i(a) T_i(t) + \beta_2 X_i(b) T_i(t) + \beta_3 X'_i(a) T_i(t) + \beta_4 X'_i(b) T_i(t)\} = g(t). \end{aligned} \quad (2.4)$$

In these equations $L \equiv p_0 d^2/dx^2 + p_1 d/dx + p_2$, $\xi_x = \partial \xi / \partial x$ and all primes, the corresponding derivatives. We next choose X_i such that

$$\begin{cases} X_1(a) = 1, & X_1(b) = X'_1(a) = X'_1(b) = 0, \\ X_2(b) = 1, & X_2(a) = X'_2(a) = X'_2(b) = 0. \end{cases} \quad (2.5)$$

and furthermore

$$\begin{cases} T_1(t) = \{\beta_2 f(t) - \alpha_2 g(t)\} / \sigma_{12}, \\ T_2(t) = \{\alpha_1 g(t) - \beta_1 f(t)\} / \sigma_{12}. \end{cases} \quad (2.6)$$

Then the boundary conditions to be satisfied by $\xi(x, t)$ are

$$\begin{cases} \alpha_1 \xi(a, t) + \alpha_2 \xi(b, t) + \alpha_3 \xi_x(a, t) + \alpha_4 \xi_x(b, t) = 0, \\ \beta_1 \xi(a, t) + \beta_2 \xi(b, t) + \beta_3 \xi_x(a, t) + \beta_4 \xi_x(b, t) = 0. \end{cases} \quad (2.7)$$

In the meantime a particular choice of $X_i(x)$ can be immediately determined by (2.5). Thus

$$X_1(x) = \frac{(x-b)^3}{(a-b)^3} - 3 \frac{(x-b)^2(x-a)}{(a-b)^3}, \quad (2.8)$$

$$X_2(x) = \frac{(x-a)^3}{(b-a)^3} - 3 \frac{(x-a)^2(x-b)}{(b-a)^3}. \quad (2.9)$$

3. With X_i , T_i determined, the right-hand side of (2.1) is known completely, and we are led to consider the following ordinary differential equation associated with (2.1),

$$L_n(\psi_n) \equiv p_0 \psi_n'' + p_1 \psi_n' + (p_2 + \lambda_n) \psi_n = 0, \quad (3.1)$$

with the boundary conditions

$$\begin{cases} U_1\{\psi_n\} = \alpha_1 \psi_n(a) + \alpha_2 \psi_n(b) + \alpha_3 \psi'_n(a) + \alpha_4 \psi'_n(b) = 0, \\ U_2\{\psi_n\} = \beta_1 \psi_n(a) + \beta_2 \psi_n(b) + \beta_3 \psi'_n(a) + \beta_4 \psi'_n(b) = 0, \end{cases} \quad (3.2)$$

where λ_n is the characteristic number. Let Ψ_{n1} , Ψ_{n2} be two fundamental solutions of (3.1), then the characteristic numbers λ_n are the roots of the following determinant:

$$\begin{vmatrix} U_1\{\Psi_{n1}\} & U_1\{\Psi_{n2}\} \\ U_2\{\Psi_{n1}\} & U_2\{\Psi_{n2}\} \end{vmatrix} = 0. \quad (3.3)$$

Now consider the following system which is adjoint to the original system defined by (3.1), (3.2):

$$L_n^*(\chi_n) \equiv L^*(\chi_n) + \lambda_n \chi_n \equiv (p_0 \chi_n)'' - (p_1 \chi_n)' + (p_2 + \lambda_n) \chi_n = 0, \quad (3.4)$$

$$\begin{cases} V_1\{\chi_n\} = \gamma_1 \chi_n(a) + \gamma_2 \chi_n(b) + \gamma_3 \chi_n'(a) + \gamma_4 \chi_n'(b) = 0, \\ V_2\{\chi_n\} = \delta_1 \chi_n(a) + \delta_2 \chi_n(b) + \delta_3 \chi_n'(a) + \delta_4 \chi_n'(b) = 0, \end{cases} \quad (3.5)$$

where $\gamma_1, \dots, \gamma_4, \delta_1, \dots, \delta_4$ are constants. The characteristic functions of the two systems are $\psi_n(x)$ and $\chi_n(x)$. It is known in the theory of differential equations [2, 3] that $\psi_n(x)$, $\chi_n(x)$ are orthogonal in the interval (a, b) . We further write $\int_a^b \psi_n \chi_n dx = C_n$. These properties will be utilized in obtaining the solution of the system (2.1) ... (2.4) in terms of expansions of $\psi_n(x)$.

We now expand the right-hand side of (2.1) into a series of $\psi_n(x)$. Let

$$X_i T'_i = T'_i \sum_{n=1}^{\infty} a_{ni} \psi_n(x), \quad (3.6)$$

$$T_i L(X_i) = T_i \sum_{n=1}^{\infty} b_{ni} \psi_n(x), \quad (3.7)$$

where

$$a_{ni} = \int_a^b X_i \chi_n dx / C_n \quad \text{etc.};$$

assuming further that

$$\xi(x, t) = \sum_{n=1}^{\infty} F_n(t) \psi_n(x), \quad (3.8)$$

and substituting (3.6), (3.7), (3.8) into (2.1), and collecting coefficients of $\psi_n(x)$ we have

$$-\lambda_n F_n(t) - F'_n(t) = \sum_{i=1}^2 \{a_{ni} T'_i - b_{ni} T_i\}, \quad (3.9)$$

where use has been made of $L(\psi_n) = -\lambda_n \psi_n$, by (3.1). Upon integration of (3.9) we have immediately

$$\begin{aligned} F_n(t) &= F_n(0) \exp \{-\lambda_n t\} \\ &\quad - \int_0^t \left\{ \sum_{i=1}^2 a_{ni} T'_i(\tau) - \sum_{i=1}^2 b_{ni} T_i(\tau) \right\} \exp \{-\lambda_n(t-\tau)\} d\tau. \end{aligned} \quad (3.10)$$

The coefficients $F_n(0)$ are to be determined by the initial condition (2.2); thus

$$\sum_{n=1}^{\infty} F_n(0) \psi_n(x) = u_0(x) - \sum_{i=1}^2 X_i(x) T_i(0),$$

and

$$F_n(0) = \int_a^b \left\{ u_0(x) - \sum_{i=1}^2 X_i(x) T_i(0) \right\} \chi_n(x) dx / C_n. \quad (3.11)$$

Integrating $\int_0^t \sum_{i=1}^2 a_{ni} T'_i(\tau) \exp \{-\lambda_n(t-\tau)\} d\tau$ in (3.10) by parts and remembering that

$$\xi(x, t) = \sum_{n=1}^{\infty} F_n(t) \psi_n(x) \quad \text{and} \quad u(x, t) = \xi(x, t) + \sum_{i=1}^2 X_i T_i,$$

we obtain the formal solution of the given system (1.1) ... (1.4) in the following form:

$$\begin{aligned} u(x, t) &= \sum_{n=1}^{\infty} \frac{\psi_n(x)}{C_n} \exp \{-\lambda_n t\} \int_a^b u_0(x) \chi_n(x) dx \\ &\quad + \sum_{i=1}^2 \sum_{n=1}^{\infty} \frac{\psi_n(x)}{C_n} \int_a^b \chi_n(x) L_n(X_i) dx \int_0^t T_i(\tau) \exp \{-\lambda_n(t-\tau)\} d\tau, \end{aligned} \quad (3.12)$$

where X_i and T_i are given by (2.8), (2.9) and (2.6).

4. As written in (3.12) the solution contains the functions X_i which are rather arbitrary: they have only to satisfy (2.5). These functions have already been determined; however, it is possible to eliminate them in the final solution. To this end we make use of the Green's formula relating the two systems (3.1), (3.2) and (3.4), (3.5)

$$\begin{aligned} &\int_a^b \{ \chi_n L_n(X_i) - X_i L_n^*(\chi_n) \} dx \\ &= p_0(b) \chi_n(b) X'_i(b) - p_0(b) \chi'_n(b) X_i(b) - p'_0(b) \chi_n(b) X_i(b) + p_1(b) \chi_n(b) X_i(b) \\ &\quad - p_0(a) \chi_n(a) X'_i(a) + p_0(a) \chi'_n(a) X_i(a) + p'_0(a) \chi_n(a) X_i(a) - p_1(a) \chi_n(a) X_i(a) \\ &= U_1 \{ X_i \} V_4 \{ \chi_n \} + U_2 \{ X_i \} V_3 \{ \chi_n \} + U_3 \{ X_i \} V_2 \{ \chi_n \} + U_4 \{ X_i \} V_1 \{ \chi_n \}, \end{aligned} \quad (4.1)$$

where $U_3 \{ X_i \}$, $U_4 \{ X_i \}$ are linear combinations of $X_i(a)$, $X_i(b)$, $X'_i(a)$, $X'_i(b)$ and $V_3 \{ \chi_n \}$, $V_4 \{ \chi_n \}$ are linear combinations of $\chi_n(a)$, $\chi_n(b)$, $\chi'_n(a)$, $\chi'_n(b)$. U_1 , U_2 and V_1 , V_2 have been defined by (3.2) and (3.5) respectively. Noting that

$$L_n^*(\chi_n) = 0, \quad V_1 \{ \chi_n \} = V_2 \{ \chi_n \} = 0,$$

and remembering $X_i(x)$ has to satisfy (2.5), we have

$$\begin{aligned} \int_a^b \chi_n L_n(X_i) dx &= \delta_{i1} \{ p_0(a) \chi'_n(a) + p'_0(a) \chi_n(a) - p_1(a) \chi_n(a) \} \\ &\quad + \delta_{i2} \{ -p_0(b) \chi'_n(b) - p'_0(b) \chi_n(b) + p_1(b) \chi_n(b) \} = \alpha_i V_4 \{ \chi_n \} + \beta_i V_3 \{ \chi_n \}, \end{aligned} \quad (4.2)$$

where δ_{i1} , δ_{i2} are the Kronecker deltas. Substituting the expressions for T_i as given by (2.6) and the result (4.2) just obtained into (3.12) we obtain the solution in the following form:

$$\begin{aligned} u(x, t) &= \sum_{n=1}^{\infty} \frac{\psi_n(x)}{C_n} \exp \{-\lambda_n t\} \int_a^b u_0(x) \chi_n(x) dx \\ &\quad + \sum_{n=1}^{\infty} \frac{\psi_n(x) V_4 \{ \chi_n \}}{C_n} \int_0^t f(\tau) \exp \{-\lambda_n(t-\tau)\} d\tau \\ &\quad + \sum_{n=1}^{\infty} \frac{\psi_n(x) V_3 \{ \chi_n \}}{C_n} \int_0^t g(\tau) \exp \{-\lambda_n(t-\tau)\} d\tau, \end{aligned} \quad (4.3)$$

where

$$V_3\{\chi_n\} = -\frac{1}{\sigma_{12}} \{ \alpha_1[(p'_0(b) - p_1(b))\chi_n(b) + p_0(b)\chi'_n(b)] \\ + \alpha_2[(p'_0(a) - p_1(a))\chi_n(a) + p_0(a)\chi'_n(a)] \}, \quad (4.4)$$

and

$$V_4\{\chi_n\} = \frac{1}{\sigma_{12}} \{ \beta_1[(p'_0(b) - p_1(b))\chi_n(b) + p_0(b)\chi'_n(b)] \\ + \beta_2[(p'_0(a) - p_1(a))\chi_n(a) + p_0(a)\chi'_n(a)] \}. \quad (4.5)$$

5. As an example of the previous discussions, consider the diffusion equation for the axi-symmetric case, $a < r < b$:

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} = \frac{\partial u}{\partial t}, \quad (5.1)$$

with the initial condition

$$u(r, t) = u_0(t), \quad t = 0, \quad (5.2)$$

and the boundary conditions

$$\begin{cases} \alpha_1 u(a, t) + \alpha_2 u(b, t) + \alpha_3 u_r(a, t) + \alpha_4 u_r(b, t) = f(t), \\ \beta_1 u(a, t) + \beta_2 u(b, t) + \beta_3 u_r(a, t) + \beta_4 u_r(b, t) = g(t). \end{cases} \quad (5.3)$$

By following the same procedure as outlined in the previous paragraphs we are led to consider the system:

$$L_n\{\psi_n\} = \left(\frac{d^2}{dr^2} + \frac{1}{r} \frac{d}{dr} + \lambda_n \right) \psi_n = 0, \quad (5.4)$$

$$\begin{cases} U_1\{\psi_n\} = \alpha_1 \psi_n(a) + \alpha_2 \psi_n(b) + \alpha_3 \psi'_n(a) + \alpha_4 \psi'_n(b) = 0, \\ U_2\{\psi_n\} = \beta_1 \psi_n(a) + \beta_2 \psi_n(b) + \beta_3 \psi'_n(a) + \beta_4 \psi'_n(b) = 0 \end{cases} \quad (5.5)$$

and the adjoint system:

$$L_n^*\{\chi_n\} \equiv \left(\frac{d^2}{dr^2} - \frac{1}{r} \frac{d}{dr} + \lambda_n + \frac{1}{r^2} \right) \chi_n = 0, \quad (5.6)$$

$$\begin{cases} V_1\{\chi_n\} = \gamma_1 \chi_n(a) + \gamma_2 \chi_n(b) + \gamma_3 \chi'_n(a) + \gamma_4 \chi'_n(b) = 0, \\ V_2\{\chi_n\} = \delta_1 \chi_n(a) + \delta_2 \chi_n(b) + \delta_3 \chi'_n(a) + \delta_4 \chi''_n(b) = 0, \end{cases} \quad (5.7)$$

where $\gamma_1, \dots, \delta_4$ are to be determined. Now the fundamental solutions of (5.4) are $J_0(\lambda_n^{1/2}r)$, $Y_0(\lambda_n^{1/2}r)$, being Bessel functions of the first and the second kind of the zero order. The characteristic function of the system (5.4), (5.5) is therefore

$$\psi_n(r) = J_0(\lambda_n^{1/2}r) - E Y_0(\lambda_n^{1/2}r), \quad (5.8)$$

where E is a constant and is determined by $U_1\{\psi_n\} = U_2\{\psi_n\} = 0$. Similarly the fundamental solutions of (5.6) are $rJ_0(\lambda_n^{1/2}r)$ and $rY_0(\lambda_n^{1/2}r)$, and the characteristic function of the system (5.6), (5.7) is

$$\chi_n(r) = rJ_0(\lambda_n^{1/2}r) - E' r Y_0(\lambda_n^{1/2}r), \quad (5.9)$$

E' to be determined by $V_1\{\chi_n\} = V_2\{\chi_n\} = 0$. The Green's formula connecting the two systems is:

$$\begin{aligned} & \int_a^b \{\chi_n L_n(\psi_n) - \psi_n L_n^*(\chi_n)\} dr \\ &= \psi'_n(b)\chi_n(b) - \psi_n(b)\chi'_n(b) + \frac{1}{b} \psi_n(b)\chi_n(b) \\ & \quad - \psi'_n(a)\chi_n(a) + \psi_n(a)\chi'_n(a) - \frac{1}{a} \psi_n(a)\chi_n(a) \\ &= U_1\{\psi_n\} V_4\{\chi_n\} + U_2\{\psi_n\} V_3\{\chi_n\} + U_3\{\psi_n\} V_2\{\chi_n\} + U_4\{\psi_n\} V_1\{\chi_n\}. \end{aligned} \quad (5.10)$$

Here $U_1\{\psi_n\}$, $U_2\{\psi_n\}$ have already been defined, as by (5.5); if we take⁴

$$U_3\{\psi_n\} = \frac{1}{\sigma_{13}\sigma_{24}} \left\{ \alpha_4 \frac{\sigma_{13}}{\sigma_{24}} \psi_n(a) + \alpha_3 \psi_n(b) \right\}, \quad (5.11)$$

$$U_4\{\psi_n\} = -\frac{1}{\sigma_{13}\sigma_{24}} \left\{ \beta_4 \frac{\sigma_{13}}{\sigma_{24}} \psi_n(a) + \beta_3 \psi_n(b) \right\}, \quad (5.12)$$

we can find $V_1\{\chi_n\}$, ..., $V_4\{\chi_n\}$ by comparing the coefficients of $\psi_n(a)$, $\psi_n(b)$, $\psi'_n(a)$, $\psi'_n(b)$ in (5.10). This results⁵

$$V_1\{\chi_n\} = \alpha_3 \sigma_{24} \left\{ \left(\frac{1}{a} - \frac{\alpha_1}{\alpha_2} \right) \chi_n(a) - \chi'_n(a) \right\} + \alpha_4 \sigma_{13} \left\{ \left(\frac{1}{b} - \frac{\alpha_2}{\alpha_4} \right) \chi_n(b) - \chi'_n(b) \right\}, \quad (5.13)$$

$$V_2\{\chi_n\} = \beta_3 \sigma_{24} \left\{ \left(\frac{1}{a} - \frac{\beta_1}{\beta_2} \right) \chi_n(a) - \chi'_n(a) \right\} + \beta_4 \sigma_{13} \left\{ \left(\frac{1}{b} - \frac{\beta_2}{\beta_4} \right) \chi_n(b) - \chi'_n(b) \right\}, \quad (5.14)$$

$$V_3\{\chi_n\} = \frac{1}{\sigma_{34}} \{ \alpha_4 \chi_n(a) + \alpha_3 \chi_n(b) \}, \quad (5.15)$$

$$V_4\{\chi_n\} = -\frac{1}{\sigma_{34}} \{ \beta_4 \chi_n(a) + \beta_3 \chi_n(b) \}. \quad (5.16)$$

Now we introduce

$$\Omega_{ij} = J_i(\lambda_n^{1/2}a) Y_j(\lambda_n^{1/2}b) - J_j(\lambda_n^{1/2}b) Y_i(\lambda_n^{1/2}a), \quad (5.17)$$

$$\begin{cases} P(\alpha) = -\alpha_1 \Omega_{01} + \alpha_3 \lambda_n^{1/2} \Omega_{11}, \\ Q(\alpha) = -\alpha_1 \Omega_{00} + \alpha_3 \lambda_n^{1/2} \Omega_{10}, \\ R(\alpha) = \alpha_2 \Omega_{00} - \alpha_4 \lambda_n^{1/2} \Omega_{01}, \\ S(\alpha) = \alpha_2 \Omega_{10} - \alpha_4 \lambda_n^{1/2} \Omega_{11}, \end{cases} \quad (5.18)$$

$$\begin{cases} K(\alpha) = \alpha_1 J_0(\lambda_n^{1/2}a) - \alpha_3 \lambda_n^{1/2} J_1(\lambda_n^{1/2}a), \\ L(\alpha) = \alpha_2 J_0(\lambda_n^{1/2}b) - \alpha_4 \lambda_n^{1/2} J_1(\lambda_n^{1/2}b), \\ M(\alpha) = \alpha_1 Y_0(\lambda_n^{1/2}a) - \alpha_3 \lambda_n^{1/2} Y_1(\lambda_n^{1/2}a), \\ N(\alpha) = \alpha_2 Y_0(\lambda_n^{1/2}b) - \alpha_4 \lambda_n^{1/2} Y_1(\lambda_n^{1/2}b), \end{cases} \quad (5.19)$$

⁴Of course, other forms of $U_3\{\psi_n\}$ and $U_4\{\psi_n\}$ may be chosen as long as $U_1\{\psi_n\}, \dots, U_4\{\psi_n\}$ form an independent set in the quantities $\psi_n(a)$, $\psi_n(b)$, $\psi'_n(a)$, $\psi'_n(b)$.

⁵Equations (5.15) and (5.16) are equivalent to (4.4) and (4.5) if we put in the latter $p_0 = 1$, $p_1(a) = 1/a$, etc., and make use of $V_1\{\chi_n\} = V_2\{\chi_n\} = 0$.

and similar expressions for $P(\beta), \dots, N(\beta)$. Also let $\sigma = b\sigma_{13}/a\sigma_{24}$, then, using the fact that $J'_0(\lambda_n^{1/2}r) = -\lambda_n^{1/2}J_1(\lambda_n^{1/2}r)$, $Y'_0(\lambda_n^{1/2}r) = -\lambda_nY_1(\lambda_n^{1/2}r)$, where the prime denotes the derivative with respect to r and that $J_1(\lambda_n^{1/2}r)Y_0(\lambda_n^{1/2}r) - Y_1(\lambda_n^{1/2}r)J_0(\lambda_n^{1/2}r) = 2/\pi\lambda_n^{1/2}r$ we obtain, after some algebraic details,

$$E = \frac{K(\alpha) + L(\alpha)}{M(\alpha) + N(\alpha)} = \frac{K(\beta) + L(\beta)}{M(\beta) + N(\beta)}, \quad (5.20)$$

$$\psi_n(a) = \frac{R(\alpha) + 2\alpha_3/\pi a}{M(\alpha) + N(\alpha)} = \frac{R(\beta) + 2\beta_3/\pi a}{M(\beta) + N(\beta)}, \quad (5.21)$$

$$\psi_n(b) = \frac{Q(\alpha) + 2\alpha_4/\pi b}{M(\alpha) + N(\alpha)} = \frac{Q(\beta) + 2\beta_4/\pi b}{M(\beta) + N(\beta)}, \quad (5.22)$$

$$E' = \frac{K(\alpha) + \sigma L(\alpha)}{M(\alpha) + \sigma N(\alpha)} = \frac{K(\beta) + \sigma L(\beta)}{M(\beta) + \sigma N(\beta)}, \quad (5.23)$$

$$\chi_n(a) = a \frac{\sigma R(\alpha) + 2\alpha_3/\pi a}{M(\alpha) + \sigma N(\alpha)} = a \frac{\sigma R(\beta) + 2\beta_3/\pi a}{M(\beta) + \sigma N(\beta)}, \quad (5.24)$$

$$\chi_n(b) = b \frac{Q(\alpha) + 2\alpha_4/\pi b}{M(\alpha) + \sigma N(\alpha)} = b \frac{Q(\beta) + 2\beta_4/\pi b}{M(\beta) + \sigma N(\beta)}. \quad (5.25)$$

Furthermore, since $\psi_m(r)$ and $\chi_n(r)$ are orthogonal in the interval (a, b) , we have

$$\int_a^b \psi_m(r)\chi_n(r) dr = 0, \quad m \neq n,$$

and

$$\begin{aligned} & \int_a^b \psi_n(r)\chi_n(r) dr \\ &= \frac{b^2}{2} \frac{\{P(\alpha) + 2\alpha_2/\pi\lambda_n^{1/2}b\}\{P(\alpha) + 2\alpha_3/\pi\lambda_n^{1/2}b\} + \{Q(\alpha) + 2\alpha_4/\pi b\}\{Q(\alpha) + 2\alpha_5/\pi b\}}{\{M(\alpha) + N(\alpha)\}\{M(\alpha) + \sigma N(\alpha)\}} \\ &- \frac{a^2}{2} \frac{\{R(\alpha) + 2\alpha_3/\pi a\}\{\sigma R(\alpha) + 2\alpha_4/\pi a\} + \{S(\alpha) + 2\alpha_1/\pi\lambda_n^{1/2}a\}\{\sigma S(\alpha) + 2\alpha_2/\pi\lambda_n^{1/2}a\}}{\{M(\alpha) + N(\alpha)\}\{M(\alpha) + \sigma N(\alpha)\}} \\ &= C_n. \end{aligned} \quad (5.26)$$

In the last expression we can replace all the $P(\alpha), \dots, N(\alpha)$ by the corresponding $P(\beta), \dots, N(\beta)$. The characteristic numbers λ_n are the roots of the equation

$$\sigma_{24}\Omega_{11}\lambda_n - (\sigma_{14}\Omega_{01} + \sigma_{32}\Omega_{10})\lambda_n^{1/2} + \sigma_{12}\Omega_{00} + \frac{2}{\pi} \left(\frac{\sigma_{13}}{a} + \frac{\sigma_{24}}{b} \right) = 0. \quad (5.27)$$

With these preliminaries the solution of the given problem can be written down according to (4.3). We note in particular that the system (5.4), (5.5) becomes self-adjoint if $a\sigma_{24} = b\sigma_{13}$, i.e., if $\sigma = 1$, in which case $E = E'$ and $\chi_n(r) = r\psi_n(r)$.

6. We now consider a special case of the previous example. Let

$$\begin{cases} \alpha_1 = -1, & \alpha_2 = 0, & \alpha_3 = \alpha, & \alpha_4 = 0, \\ \beta_1 = 0, & \beta_2 = 1, & \beta_3 = 0, & \beta_4 = \beta, \end{cases} \quad (6.1)$$

then the boundary conditions are reduced to

$$\begin{cases} -u(a, t) + \alpha u_r(a, t) = f(t), \\ u(b, t) + \beta u_r(b, t) = g(t). \end{cases} \quad (6.2)$$

This problem arises in the evaluation of transient temperature distribution in a homogeneous hollow circular cylinder, $a \leq r \leq b$, when the gas temperatures inside and outside the cylinder are functions of time, being $f(t)$ and $g(t)$ respectively, and the initial temperature distribution in the cylinder is $u_0(r)$. The constants α and β are associated with the heat transfer coefficient and the thermal conductivity of the cylinder material (here $\alpha, \beta > 0$). Then $\alpha\sigma_{24} = \beta\sigma_{13} = 0$, and the system is self-adjoint. Here we have

$$E = E' = \frac{J_0(\lambda_n^{1/2}a) + \alpha\lambda_n^{1/2}J_1(\lambda_n^{1/2}a)}{Y_0(\lambda_n^{1/2}a) + \alpha\lambda_n^{1/2}Y_1(\lambda_n^{1/2}a)} = \frac{J_0(\lambda_n^{1/2}b) - \beta\lambda_n^{1/2}J_1(\lambda_n^{1/2}b)}{Y_0(\lambda_n^{1/2}b) - \beta\lambda_n^{1/2}Y_1(\lambda_n^{1/2}b)}, \quad (6.3)$$

$$\psi_n(r) = \frac{1}{r}\chi_n(r) = J_0(\lambda_n^{1/2}r) - E Y_0(\lambda_n^{1/2}r) \quad (6.4)$$

$$\psi_n(a) = \frac{1}{a}\chi_n(a) = -\frac{2\alpha/\pi a}{Y_0(\lambda_n^{1/2}a) + \alpha\lambda_n^{1/2}Y_1(\lambda_n^{1/2}a)} = \frac{\Omega_{00} - \beta\lambda_n^{1/2}\Omega_{01}}{Y_0(\lambda_n^{1/2}b) - \beta\lambda_n^{1/2}Y_1(\lambda_n^{1/2}b)} \quad (6.5)$$

$$\psi_n(b) = \frac{1}{b}\chi_n(b) = -\frac{\Omega_{00} + \alpha\lambda_n^{1/2}\Omega_{10}}{Y_0(\lambda_n^{1/2}b) + \alpha\lambda_n^{1/2}Y_1(\lambda_n^{1/2}b)} = \frac{2\beta/\pi b}{Y_0(\lambda_n^{1/2}b) - \beta\lambda_n^{1/2}Y_1(\lambda_n^{1/2}b)} \quad (6.6)$$

$$\begin{aligned} C_n &= \int_a^b \psi_n(r)\chi_n(r) dr = \int_a^b r\psi_n'(r) dr \\ &= \frac{2}{\pi^2\lambda_n} \left\{ \frac{1 + \beta^2\lambda_n}{\{Y_0(\lambda_n^{1/2}b) - \beta\lambda_n^{1/2}Y_1(\lambda_n^{1/2}b)\}^2} - \frac{1 + \alpha^2\lambda_n}{\{Y_0(\lambda_n^{1/2}a) + \alpha\lambda_n^{1/2}Y_1(\lambda_n^{1/2}a)\}^2} \right\}, \end{aligned} \quad (6.7)$$

and λ_n is a root of

$$\alpha\beta\Omega_{11}\lambda_n - (\alpha\Omega_{10} - \beta\Omega_{01})\lambda_n^{1/2} - \Omega_{00} = 0. \quad (6.8)$$

Furthermore, from (5.15) and (5.16) we have

$$V_3\{\chi_n\} = \chi_n(b)/\beta,$$

$$V_4\{\chi_n\} = -\chi_n(a)/\alpha,$$

so that the solution of this special case can be written down, according to (4.3),

$$\begin{aligned} u(r, t) &= \sum_{n=1}^{\infty} \frac{\psi_n(r)}{C_n} \exp(-\lambda_n t) \int_a^b r u_0(r) \psi_n(r) dr \\ &\quad - \frac{a}{\alpha} \sum_{n=1}^{\infty} \frac{\psi_n(r)\psi_n(a)}{C_n} \int_0^t f(\tau) \exp\{-\lambda_n(t-\tau)\} d\tau \\ &\quad + \frac{b}{\beta} \sum_{n=1}^{\infty} \frac{\psi_n(r)\psi_n(b)}{C_n} \int_0^t g(\tau) \exp\{-\lambda_n(t-\tau)\} d\tau. \end{aligned} \quad (6.9)$$

The solution of this special case for $f(t)$ and $g(t)$ equal to constants has already been given by Carslaw and Jaeger [4], using the method of the Laplace transform. It is easy to verify that the result presented here is the same as that given in [4]. For, if $\varphi(r) =$

$\sum_{n=1}^{\infty} A_n \psi_n(r)$, $a < r < b$, then

$$\begin{aligned} A_n &= \frac{1}{C_n} \int_a^b r\varphi(r) \psi_n(r) dr = -\frac{1}{C_n \lambda_n} \int_a^b r\varphi(r) \left(\psi_n'' + \frac{1}{r} \psi_n' \right) dr \\ &= \frac{1}{C_n \lambda_n} \left\{ \frac{b}{\beta} \varphi(b) \psi_n(b) + \frac{a}{\alpha} \varphi(a) \psi_n(a) + \int_a^b r\varphi'(r) \psi_n'(r) dr \right\}, \end{aligned} \quad (6.10)$$

where use has been made of $-\psi_n(a) + \alpha\psi_n'(a) = \psi_n(b) + \beta\psi_n'(b) = 0$. By putting $\varphi(r) = 1$ and $\log r$ successively in (6.10) we can find the summation of the series $\sum_{n=1}^{\infty} \psi_n(r) \psi_n(a)/C_n \lambda_n$ and $\sum_{n=1}^{\infty} \psi_n(r) \psi_n(b)/C_n \lambda_n$. Replacing $f(t)$ and $g(t)$ by the constants f and g respectively in (6.9) and using the results just obtained after performing the integration, we get

$$\begin{aligned} u(r, t) &= \sum_{n=1}^{\infty} \frac{\psi_n(r)}{C_n} \exp(-\lambda_n t) \int_a^b r u_0(r) \psi_n(r) dr \\ &\quad - \frac{af\{b \log b/r + \beta\} + bg\{a \log a/r - \alpha\}}{ab \log b/a + a\beta + b\alpha} \\ &\quad + \sum_{n=1}^{\infty} \frac{\psi_n(r)}{C_n \lambda_n} \left\{ \frac{af}{\alpha} \psi_n(a) - \frac{bg}{\beta} \psi_n(b) \right\} \exp(-\lambda_n t), \end{aligned} \quad (6.11)$$

which is the form given by Carslaw and Jaeger. In this expression $\psi_n(r)$, $\psi_n(a)$ and $\psi_n(b)$ are given by (6.4), (6.5) and (6.6) respectively.

REFERENCES

1. Courant and Hilbert, *Methods mathematical physics*, Vol. 1 (Interscience Publishers, 1953), 277.
2. E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, 1955
3. E. L. Ince, *Ordinary differential equations*, Dover, Chapters IX and X
4. H. S. Carslaw and J. C. Jaeger, *Conduction of heat in solids*, Oxford University Press, pp. 278-279, 1950

BOOK REVIEWS

Theorie des Circuits de Telecommunication. By Vitold Belevitch. Librairie Universitaire Louvain, 1957. viii 384 pp. \$9.00.

This is a well conceived exposition of electric circuit theory for the use of engineers having the usual mathematical preparation. The contents of the book are indicated by the titles of the fourteen chapters, which can be translated freely as follows: I, Linear systems; II, Analysis of Kirchoff networks; III, Energetics of passive networks; IV, Reflection and transmission; V, Scattering matrix; VI, Image parameters; VII, Analytic theory of passive networks; VIII, Image theory of low-pass filters; IX, Bandpass filters; X, Filters with pre-determined attenuations; XI, Inductances and transformers; XII, Amplifiers; XIII, Transient phenomena; XIV, Complements on the synthesis of passive networks.

Thus the book is devoted chiefly to the *steady state* theory of circuits consisting of discrete elements. It appears that the author's knowledge of this subject is very extensive, and that he has made many significant contributions to it; and, so far as the reviewer can judge, he expounds the subject with great competence and thoroughness. On the other hand, the parts of the book (notably Chapter XIII) which deal with other branches of circuit theory seem to be too sketchy to serve as more than rudimentary introductions.

As has been indicated above, the mathematical preparation expected of the reader is quite modest. Not only does the mathematics used not go appreciably beyond that usually taught in engineering curricula, but also in many places intuitive physical arguments are used, instead of rigorous mathematical demonstrations, in the derivation of formulae. On the logical level which the author has chosen, the exposition is generally satisfactory. However, there are occasional slips which may give unwary readers trouble. Thus the author states the Hurwitz stability condition (i.e. the condition that the real parts of the roots of an algebraic equation be negative) in a special simplified form which applies only to the case in which the leading coefficient in the equation is positive, but he does not mention this restrictive assumption concerning the coefficients. Again, in the discussion of Fourier integrals, the author states that the absolute integrability of $f(t)$ over the interval $(-\infty, \infty)$ is a *necessary* condition for the representability of $f(t)$ by a Fourier integral.

There is an extensive list of errata. However, this is concerned mainly with trivial typographical errors, and does not seem to touch on any of the more serious faults.

On the whole, the book affords a comprehensive and lucid exposition of a large and important part of circuit theory, and it should be welcome to engineers who are concerned with the practical design and use of circuits. An excellent feature of the book is a fifteen-page critical bibliography. This will afford adequate guidance to readers who wish to study the subject more deeply and extensively.

L. A. MACCOLL

Wahrscheinlichkeitstheorie. By Hans Richter. Springer-Verlag, Berlin, Göttingen, Heidelberg, 1956. xi 435 pp. \$16.65.

The author states in the preface that his objective is to supply a remedy for the lack of any modern German text on probability theory. He seems to have accomplished not only this modest goal but also to have made a valuable contribution to the international literature.

The reader is assumed to have some knowledge of analysis and linear algebra plus some appreciation of pure mathematics. Starting at this relatively low level, the author devotes about half the book to a rigorous development of the experimental and mathematical foundations including the necessary background from the theory of measure and integration. The last half of the book includes discussions of random variables (expectations, etc.), properties of common types of distributions (Poisson, Gaussian, etc.) and finally some of the well-known limit theorems (laws of large numbers, zero-one laws, central limit theorems, etc.).

Although the book covers only the above limited range of topics, it treats these very thoroughly. There is little if any discussion of applications or more advanced topics such as Markov chains or stochastic processes.

G. F. NEWELL

(Continued on p. 258)

TOROIDAL WAVE FUNCTIONS*

BY

V. H. WESTON

University of Toronto

Abstract. The Helmholtz equation is solved in toroidal coordinates. A complete set of solutions is obtained representing radiations from a ring source.

Introduction. Up to now the Helmholtz equation has been solved only for separable coordinate systems. This paper presents solutions for a non-separable coordinate system, the toroidal. The main interest will be with continuous, single-valued solutions satisfying the radiation condition and possessing a ring singularity. Exact expressions for each of the wave functions will be obtained in the form of a series expansion and integral over finite range. The series expansions are uniformly convergent everywhere in space except, of course, at the ring singularity.

The time dependence will be of the form $\exp(-i\omega t)$.

1. Series solution of the Helmholtz equation in toroidal coordinates. The relation between toroidal and cartesian coordinates systems is given by [5, p. 151]

$$\begin{aligned} x &= \frac{d \sinh \xi \cos \phi}{\cosh \xi - \cos \eta}, \\ y &= \frac{d \sinh \xi \sin \phi}{\cosh \xi - \cos \eta}, \\ z &= \frac{d \sin \eta}{\cosh \xi - \cos \eta}. \end{aligned} \quad (1.1)$$

Domains of the coordinates are $0 \leq \eta \leq 2\pi$, $0 \leq \phi \leq 2\pi$, $0 \leq \xi \leq \infty$ where $\xi = \xi_0$ defines a torus

$$z^2 + (\rho - d \coth \xi_0)^2 = d^2 \operatorname{esch}^2 \xi_0$$

and $\eta = \eta_0$ defines a sphere

$$(z - d \cot \eta_0)^2 + \rho^2 = d^2 \csc^2 \eta_0$$

where $\rho = (x^2 + y^2)^{1/2}$.

The metric coefficients are given by the following relations

$$\begin{aligned} h_\xi &= h_\eta = \frac{d}{\cosh \xi - \cos \eta}, \\ h_\phi &= \frac{d \sinh \xi}{\cosh \xi - \cos \eta}. \end{aligned} \quad (1.2)$$

From now on, in order to facilitate analysis, the variable s will be used instead of ξ where the two are related by the equation $\cosh \xi = s$.

Now it has been shown [7], that for a certain class of non-separable rotational coordinates (u_1, u_2, ϕ) , there are solutions of

$$\nabla^2 \psi + k^2 \psi = 0 \quad (1.3)$$

*Received May 1, 1957; revised manuscript received July 2, 1957.

given by

$$\begin{aligned}\psi(u_1, u_2, \phi) &= e^{i\mu\phi} \sum_r a_r(u_i)[h_3(u_1, u_2)]^r \\ \psi(u_1, u_2, \phi) &= e^{i\mu\phi} B(u_i) \sum_r b_r(u_i)[h_3(u_1, u_2)]^r\end{aligned}\quad (1.4)$$

where $i \neq 3$, and $i \neq j \neq 3$, provided that the metric coefficients h_3 and h_1 , and the function $B(u_i)$ satisfy certain conditions. The coefficients $a_r(u_i)$ and $b_r(u_i)$ of the above series satisfy a recurrence set of ordinary differential equations.

In particular, the toroidal coordinates (ξ, η, ϕ) belong to this rotational class and the solutions are such that the power series given by (1.4) have a lower termination. If the power series are expressed in the variable $(s - \cos \eta)^{-1}$ instead of h_ϕ the differential equations involving the coefficients of the power series take a simpler form. The function $B(\eta)$ for toroidal coordinates is $\sin \eta$. Solutions of (1.3) are given by

$$\psi_e(s, \eta, \phi) = e^{i\mu\phi} \sum_{r=T}^{\infty} A_r(s) (s - \cos \eta)^{-r} \quad (1.5)$$

and

$$\psi_0(s, \eta, \phi) = e^{i\mu\phi} \sin \eta \sum_{r=T'}^{\infty} B_r(s) (s - \cos \eta)^{-r}. \quad (1.6)$$

The coefficients must satisfy the differential equations

$$\begin{aligned}(s^2 - 1)A_r'' + 2sA_r' - A_r \left[r(r+1) + \frac{\mu^2}{s^2 - 1} \right] + k^2 d^2 A_{r-2} \\ - (2r-1)[(s^2 - 1)A_{r-1}' - s(r-1)A_{r-1}] = 0,\end{aligned}\quad (1.7)$$

$$\begin{aligned}(s^2 - 1)B_r'' + 2sB_r' - B_r \left[r(r-1) + \frac{\mu^2}{s^2 - 1} \right] + k^2 d^2 B_{r-2} \\ - (2r-1)[(s^2 - 1)B_{r-1}' - s(r-2)B_{r-1}] = 0,\end{aligned}\quad (1.8)$$

where the prime denotes differentiation with respect to s . The numbers T and T' are determined from the boundary conditions. The problem of solving a partial differential equation in three variables in which only one variable is separable is reduced to solving a recurrence set of ordinary differential equations in one variable.

The homogeneous equation corresponding to the equation (1.7) is

$$(s^2 - 1)\omega'' + 2s\omega' - r(r+1)\omega - \frac{\mu^2}{s^2 - 1}\omega = 0 \quad (1.9)$$

of which, the associated Legendre functions $P_r^\mu(s)$ and $Q_r^\mu(s)$ are solutions. The function $\omega_r^\mu(s)$ will be used to represent both solutions of (1.9).

It is immediately evident that for $r = T$ in (1.7) and $r = T'$ in (1.8), the non-homogeneous portion of the equations vanish. Hence one has $A_T = \omega_T^\mu(s)$ and $B_{T'} = \omega_{T'-1}^\mu(s)$. Since the solutions of the corresponding homogeneous equations of (1.7) and (1.8) are known, the equations can be solved easily, resulting in the following relations,

$$A_r(s) = \omega_r^\mu(s) \int_{c_1}^s \frac{dz}{(1-z^2)[\omega_r^\mu(z)]^2} \int_{c_2}^z F_r(x)\omega_r^\mu(x) dx \quad (1.10)$$

and

$$B_r(s) = \omega_{r-1}^\mu(s) \int_{c_1}^s \frac{dz}{(1-z^2)[\omega_{r-1}^\mu(z)]^2} \int_{c_1}^s G_r(x) \omega_{r-1}^\mu(x) dx, \quad (1.11)$$

where

$$F_r(s) = k^2 d^2 A_{r-2} - (2r-1)[(s^2-1)A'_{r-1} - s(r-1)A_{r-1}] \quad (1.12)$$

$$G_r(s) = k^2 d^2 B_{r-2} - (2r-1)[(s^2-1)B'_{r-1} - s(r-2)B_{r-1}]. \quad (1.13)$$

The constants c_1 , c_2 , c'_1 and c'_2 are determined from the boundary conditions.

Define a basic set of solutions $\psi_{eT}^\mu(P)$ and $\psi_{eT}^\mu(Q)$ such that $A_T(s)$ is equal to $P_T^\mu(s)$ and $Q_T^\mu(s)$ respectively and the constants of integration in the expression for $A_r(s)$ are taken as fixed numbers.

Let $\psi_{eT}^\mu(s, \eta, \phi)$ be a solution of Eq. (1.3) of the form (1.5) with arbitrary constants of integration and with $A_T(s) = a_1 P_T^\mu(s) + b_1 Q_T^\mu(s)$. Then

$$\psi_{eT}^\mu(s, \eta, \phi) = a_1 \psi_{eT}^\mu(P) + b_1 \psi_{eT}^\mu(Q)$$

is a solution of Eq. (1.3), with the coefficient of $(s - \cos \eta)^{-T}$ vanishing. Thus we have

$$\psi_{eT}^\mu(s, \eta, \phi) = a_1 \psi_{eT}^\mu(P) + b_1 \psi_{eT}^\mu(Q) = \psi_{eT+1}^\mu(s, \eta, \phi),$$

where $\psi_{eT+1}^\mu(s, \eta, \phi)$ is a solution of Eq. (1.3) of the form (1.5) where the lower limit of summation is $T+1$. The coefficient of $(s - \cos \eta)^{-T-1}$ in $\psi_{eT+1}^\mu(s, \eta, \phi)$ must be of the form $a_2 P_{T+1}^\mu(s) + b_2 Q_{T+1}^\mu(s)$. Hence in a similar manner we have

$$\psi_{eT}^\mu(s, \eta, \phi) = a_1 \psi_{eT}^\mu(P) + a_2 \psi_{eT+1}^\mu(P) + b_1 \psi_{eT}^\mu(Q) + b_2 \psi_{eT+1}^\mu(Q) = \psi_{eT+2}^\mu(s, \eta, \phi).$$

By mathematical induction we see that any solution $\psi_{eT}^\mu(s, \eta, \phi)$ with arbitrary constants of integration in the expressions for $A_r(s)$, is just a linear combination of $\psi_{ep}^\mu(P)$ and $\psi_{ep}^\mu(Q)$ where $p = T, T+1, T+2, \dots$. A similar discussion follows for the solutions of the form $\psi_{eT'}^\mu(s, \eta, \phi)$. So no restriction is placed if we take the constants in the integrals (1.10) and (1.11) as pre-determined fixed numbers, thus allowing us to obtain explicit expressions for a set of solutions of (1.3).

Any solutions of the form $\{\psi_{eT}^\mu(s, \eta, \phi) + \psi_{eT'}^\mu(s, \eta, \phi)\}$ with arbitrary constants of integration can be expressed as a linear combination of the explicit solutions $\psi_{ep}^\mu(P)$, $\psi_{ep}^\mu(Q)$, $\psi_{op}^\mu(P)$ and $\psi_{op}^\mu(Q)$ where $p = T, T+1, \dots$ and $p' = T', T'+1, \dots$.

From now on we shall be interested in solutions periodic in the angle ϕ . Hence we set $\mu = m$ where $m = 0, \pm 1, \pm 2, \dots$

2. Obtaining explicit expressions for $\psi_{eT}^m(P)$ and $\psi_{eT'}^m(P)$. The wave functions $\psi_{eT}^m(P)$ and $\psi_{eT'}^m(P)$ are defined as those solutions satisfying (1.5) and (1.6) respectively where the coefficients $A_r(s)$ and $B_r(s)$ are given by (1.10) and (1.11) with the constants c_1, c_2, c'_1, c'_2 all set equal to unity, and $A_T(s)$ and $B_{T'}(s)$ equal to $P_T^{|m|}(s)$ and $P_{T'-1}^{|m|}(s)$ respectively. The negative superscript is taken so that our results include the case in which T and $T'-1$ are integers such that $|T| < |m|$ and $|T'-1| < |m|$ when the associated Legendre functions $P_T^{|m|}(s)$ and $P_{T'-1}^{|m|}(s)$ do not exist.

For convenience the symbol M will be used to signify $|m|$, where m is a positive or negative integer or zero, i.e.

$$M = |m| = 0, 1, 2, 3, \dots \quad (2.1)$$

The integral operator $I(M, r)$ operating on the function $F(s)$ is defined as follows:

$$I(M, r)F(s) = \omega_r^M(s) \int_1^s \frac{dz}{(1-z^2)[\omega_r^M(z)]^2} \int_1^z \omega_r^M(x)F(x) dx. \quad (2.2)$$

Hence the coefficients $A_r(s)$ and $B_r(s)$ for $\psi_{rT}^m(P)$ and $\psi_{0T}^m(P)$ satisfy the relations

$$A_r(s) = I(M, r)F_r(s), \quad (2.3)$$

$$B_r(s) = I(M, r-1)G_r(s), \quad (2.4)$$

where $F_r(s)$ and $G_r(s)$ are defined by (1.12) and (1.13). To calculate $A_r(s)$ and $B_r(s)$ the following lemma* is needed:

Lemma: If

$$G(M, X, X-R) = (s^2 - 1)^{X/2} P_{X-R-1}^{-M-X}(s), \quad (2.5)$$

where X is a positive, and M a non-negative integer and if

$$F(s) = a \left[(s^2 - 1) \frac{d}{dx} G(M, X-1, X-R) - s(R-1)G(M, X-1, X-R) \right] + bG(M, X-1, X-R+1), \quad (2.6)$$

then

$$I(M, R)F(s) = -[a(M+2X-R-1) + b]G(M, X, X-R)[2X]^{-1}. \quad (2.7)$$

Now $A_T(s) = P_T^{-M}(s) = P_{-T-1}^{-M}(s) = G(M, 0, -T)$ hence setting $r = T+1$ in (1.13) and using the fact that $A_{T-1}(s) \equiv 0$, one has

$$F_{T+1}(s) = -(2T+1) \left[(s^2 - 1) \frac{d}{ds} G(M, 0, -T) - sTG(M, 0, -T) \right]. \quad (2.8)$$

But

$$A_{T+1}(s) = I(M, T+1)F_{T+1}(s). \quad (2.9)$$

Now $F_{T+1}(s)$ corresponds to expression (2.6) where

$$a = -(2T+1), \quad b = 0, \quad X = 1, \quad R = T+1,$$

hence the lemma gives A_{T+1} from (2.9) and (2.7), resulting in the following relation

$$\begin{aligned} A_{T+1} &= -2^{-1}[-(2T+1)(M-T)]G(M, 1, -T) \\ &= (T+\frac{1}{2})(M-T)(s^2 - 1)^{1/2} P_{-T-1}^{-M-1}(s). \end{aligned} \quad (2.10)$$

To obtain $A_{T+2}(s)$ one must break the operation

$$A_{T+2}(s) = I(M, T+2)F_{T+2}(s) \quad (2.11)$$

into two separate integrals, i.e. $F_{T+2}(s)$ must be broken up into the following two expressions

$$-(2T+3)[(s^2 - 1)A'_{T+1} - s(T+1)A_{T+1}] \quad (2.12)$$

*This is a consequence of Lemma 2, page 13, [6].

and

$$k^2 d^2 A_T . \quad (2.13)$$

Expression (2.12) reduces to the form given by (2.6) where

$$a = -(2T + 3)(T + \frac{1}{2})(M - T), \quad b = 0, \quad X = 2, \quad R = T + 2$$

and (2.13) reduces to the form given by (2.6) where

$$a = 0, \quad b = k^2 d^2, \quad X = 1, \quad R = T + 2.$$

Hence using (2.11) and (2.7) one obtains

$$\begin{aligned} A_{T+2} &= -[-(2T + 3)(T + \frac{1}{2})(M - T)(M - T + 1)]G(M, 2, -T)4^{-1} \\ &\quad - k^2 d^2 G(M, 1, -T - 1)2^{-1} \\ &= \frac{1}{2}(T + 3/2)(T + \frac{1}{2})(M - T)(M - T + 1)(s^2 - 1)P_{-T-1}^{-M-2}(s) \\ &\quad - \frac{k^2 d^2}{2}(s^2 - 1)^{1/2}P_{-T-2}^{-M-1}(s). \end{aligned} \quad (2.14)$$

Rather than calculate the remaining $A_r(s)$ in a similar manner it is better to obtain them through mathematical induction. Assume that

$$A_r(s) = \sum_{t=0}^{\lfloor(r-T)/2\rfloor} (k d)^{2t} A'_r G(M, r - T - t, -T - t), \quad (2.15)$$

where A'_r are constants. In the expression for the upper limit of summation, the following notation is used: $[x]$ is the integer such that $x - 1 < [x] \leq x$. The above expression obviously holds for $r = T, T + 1$ and $T + 2$.

Assume that the expression holds for $r - 1$ and $r - 2$, hence in order for it to hold for r one must have

$$A_r(s) = I(M, r)F_r(s), \quad (2.16)$$

where from (1.13) and (2.15), $F_r(s)$ becomes

$$\begin{aligned} F_r(s) &= k^2 d^2 \sum_{t=0}^{\lfloor(r-T-2)/2\rfloor} (k d)^{2t} A'_{r-2} G(M, r - 2 - T - t, -T - t) \\ &\quad - (2r - 1) \left[(s^2 - 1) \frac{d}{ds} - s(r - 1) \right] \\ &\quad \sum_{t=0}^{\lfloor(r-T-1)/2\rfloor} (k d)^{2t} A'_{r-1} G(M, r - 1 - T - t, -T - t). \end{aligned}$$

The coefficient of $(kd)^{2t}$ in $F_r(s)$ is

$$\begin{aligned} A'_{r-2} G(M, r - 1 - T - t, -T - t + 1) \\ - (2r - 1) \left[(s^2 - 1) \frac{d}{ds} - s(r - 1) \right] A'_{r-1} G(M, r - 1 - T - t, -T - t). \end{aligned}$$

Hence the coefficient of $(kd)^{2t}$ in $I(M, r)F_r(s)$ is, on setting $a = -(2r - 1)A'_{r-1}$, $b = A'_{r-2}$, $X = r - T - t$, $R = r$ in expression (2.6) and using (2.7)

$$[(M - 2T - 2t + r - 1)(2r - 1)A'_{r-1} - A'_{r-2}] \frac{G(M, r - T - t, -T - t)}{2(r - T - t)}. \quad (2.17)$$

Equating coefficients of $(kd)^{2t}$ in (2.16) one obtains the relation

$$\begin{aligned} A_r^t G(M, r - T - t, -T - t) \\ = [(M - 2T - 2t + r - 1)(2r - 1)A_{r-1}^t \\ - A_{r-2}^{t-1}] \frac{G(M, r - T - t, -T - t)}{2(r - T - t)} \end{aligned} \quad (2.18)$$

which shows that the expression (2.15) holds for r if it holds for $r - 1$ and $r - 2$, provided that the constants A_r^t satisfy the equality

$$A_r^t = [(M - 2t - 2T + r - 1)(2r - 1)A_{r-1}^t - A_{r-2}^{t-1}] 2^{-1}(r - T - t)^{-1}. \quad (2.19)$$

Thus one obtains

$$A_r^t = \frac{(-1)^{t+r-T} \Gamma(-T - t + \frac{1}{2}) \Gamma(-2T + r - 2t + M)}{2^t (r - T - 2t)! (t)! \Gamma(-r + t + \frac{1}{2}) \Gamma(M - T)}. \quad (2.20)$$

This expression for the constants A_r^t holds for $r = T, T + 1, T + 2$ as is seen when comparing the values given by (2.20) for the cases $r = T + 1, t = 0; r = T + 2, t = 0$; and $r = T + 2, t = 1$ with values of constants in equations (2.10) and (2.14). Since the expression (2.15) holds for $r = T, T + 1, T + 2$ and it has been shown that if it holds for $r - 1, r - 2$ then it will hold for r , one can, by mathematical induction, conclude that (2.15) holds for every r .

Hence one can immediately write

$$\psi_{eT}^m(P) = e^{im\phi} \sum_{r=T}^{\infty} (s - \cos \eta)^{-r} \sum_{t=0}^{\lfloor (r-T)/2 \rfloor} (k d)^{2t} A_r^t (s^2 - 1)^{(r-T-t)/2} P_{-T-t-1}^{-M-r+T+t}(s). \quad (2.21)$$

Noting that $B_{T'}(s) = P_{-T'}^{-M}(s)$ we can obtain in a manner similar to the above the following explicit expression for $\psi_{0T'}^m(P)$

$$\begin{aligned} \psi_{0T'}^m(P) = e^{im\phi} \sin \eta \sum_{r=T'}^{\infty} (s - \cos \eta)^{-r} \sum_{t=0}^{\lfloor (r-T')/2 \rfloor} (k d)^{2t} \\ \cdot B_r^t (s^2 - 1)^{(r-T'-t)/2} P_{-T'-t}^{-M-r+T'+t}(s), \end{aligned} \quad (2.22)$$

where the constants B_r^t are given by

$$B_r^t = \frac{(-1)^{t+r-T'} \Gamma(-T' - t + .5) \Gamma(-2T' + r - 2t + M + 1)}{2^t (t)! (r - T' - 2t)! \Gamma(-r + t + .5) \Gamma(M - T' + 1)}. \quad (2.23)$$

The expressions (2.21) and (2.22) can be placed in a neater form. In (2.21), interchange the order of summation of r and t , obtaining

$$\sum_{r=T}^{\infty} \sum_{t=0}^{\lfloor (r-T)/2 \rfloor} = \sum_{t=0}^{\infty} \sum_{r=2t+T}^{\infty}.$$

(The validity of this operation will be shown below.) Then replace the summation over r by summation over σ where $\sigma = -2t + (r - T)$. One then has

$$\psi_{eT}^m(P) = e^{im\phi} \sum_{t=0}^{\infty} \sum_{\sigma=0}^{\infty} \frac{(k d)^{2t} a_{\sigma}^t (s^2 - 1)^{(\sigma+t)/2}}{(s - \cos \eta)^{2t+\sigma+T}} P_{T+t}^{-M-t-\sigma}(s), \quad (2.24)$$

where

$$a_\sigma^t = \frac{(-1)^t(M - T)_\sigma(T + t + .5)_\sigma}{2^t(t)!(\sigma)!}. \quad (2.25)$$

Similarly $\psi_{0T'}^m(P)$ reduces to the form

$$\psi_{0T'}^m(P) = e^{im\phi} \sin \eta \sum_{t=0}^{\infty} \sum_{\sigma=0}^{\infty} \frac{(k d)^{2t} b_\sigma^t(s^2 - 1)^{(\sigma+t)/2}}{(s - \cos \eta)^{2t+\sigma+T'}} P_{T'+t-1}^{-M-t-\sigma}(s), \quad (2.26)$$

where

$$b_\sigma^t = \frac{(-1)^t(M + 1 - T')_\sigma(T' + t + .5)_\sigma}{2^t(t)!(\sigma)!}. \quad (2.27)$$

To show convergence of the above series we require the following inequality

$$P_r^{-M}(s) \leq (s - 1)^{M/2}(s + 1)^{-M/2}[s \pm (s^2 - 1)^{1/2}]^r / \Gamma(1 + M), \quad (2.28)$$

where the positive and negative signs are taken when $r \geq 0$ and $0 > r$ respectively. We shall represent the series expansion in (2.24) by the following

$$\psi_e^m(P) = e^{im\phi} \sum_{t=0}^{\infty} \sum_{\sigma=0}^{\infty} C_{t\sigma}(s, \eta). \quad (2.29)$$

Using (2.28) and the inequality

$$(s - 1)(s - \cos \eta)^{-1} \leq 1$$

which holds for $1 \leq s \leq \infty$ and $0 \leq \eta \leq 2\pi$ we can show that the series $\sum_{\sigma=0}^{\infty} |C_{t\sigma}(s, \eta)|$ is dominated by the absolutely convergent series $D_t(s, \eta) \sum_{\sigma=0}^{\infty} |d_{t\sigma}|$ where

$$|d_{t\sigma}| = \left| \frac{(M - T)_\sigma(T + t + .5)_\sigma}{(\sigma)!(1 + M + t)_\sigma} \right| \quad (2.30)$$

and

$$D_t(s, \eta) = \left(\frac{s - 1}{s + 1} \right)^{M/2} \frac{k^{2t} d^{2t} (s - 1)^t [s \pm (s^2 - 1)^{1/2}]^{t+T}}{2^t(t)!\Gamma(1 + M + t)(s - \cos \eta)^{2t+T}},$$

where the positive sign is taken when $t + T \geq 0$, negative sign otherwise.

For large t , ($t + T + \frac{1}{2} > 0$), we have

$$\sum_{\sigma=0}^{\infty} |d_{t\sigma}| \leq \frac{\Gamma(1 + M + t)(\pi)^{1/2}}{\Gamma(1 + t + T) \Gamma(.5 + M - T)} + K \left[1 + O\left(\frac{1}{t}\right) \right], \quad (2.31)$$

where K is a constant which vanishes if $M \geq T$.

Using (2.31) it can be shown that the series $\sum_{t=0}^{\infty} \sum_{\sigma=0}^{\infty} D_t(s, \eta) |d_{t\sigma}|$ is absolutely convergent for every value of $k d$ and every value of s and η in the ranges $0 \leq s \leq \infty$ and $0 \leq \eta \leq 2\pi$. Since the series $\sum \sum |c_{t\sigma}|$ is dominated by $\sum \sum D_t(s, \eta) |d_{t\sigma}|$ we can say that the original series given by (2.24) is uniformly and absolutely convergent for $1 \leq s \leq \infty$ and $0 \leq \eta \leq 2\pi$. Hence the change of order of summation above is valid.

A similar discussion follows for the series (2.26) except that the region of uniform convergence $1 \leq s \leq \infty$ and $0 \leq \eta \leq 2\pi$, holds only if $T' - 1 - M$ is a non-negative integer. For other values of $T' - 1 - M$ the region of uniform convergence is $1 \leq s \leq \infty$ and $0 < \eta < 2\pi$, the wave functions $\psi_{0T'}^m(P)$ being discontinuous along the lines $\eta = 0$ and $\eta = 2\pi$.

3. Determination of T , T' through continuity conditions. We will be concerned with solutions of the Helmholtz equation for the exterior problem, that is, solutions which are non-singular and continuous in the region external to the torus $s = s_0$, where by external we mean $s_0 \geq s \geq 1$. The region $s_0 \leq s \leq \infty$ describes the surface and interior of a torus. The limiting torus $s = \infty$ describes a ring with radius d and centre, the origin. As is seen from (1.1), a portion of the surface of the torus $s = 1$ is the z -axis. The rest of the surface extends throughout infinity in all directions, i.e. if R is the spherical polar coordinate, then $R = \infty$ for torus $s = 1$.

We shall eventually consider solutions satisfying the radiation condition, but before doing so, we must consider the effect of applying the conditions of continuity and non-singularity in the region $s_0 \geq s \geq 1$. Since $Q_T^m(s)$ is singular on the surface $s = 1$, solutions involving the associated Legendre functions $Q_T^m(s)$ will be singular there. Hence it is seen that solutions which are non-singular in region $s_0 \geq s \geq 1$ can be formed only from $\psi_{eT}^m(P)$ and $\psi_{oT'}^m(P)$ solutions. The condition of continuity shall be applied to the $\psi_{eT}^m(P)$ and $\psi_{oT'}^m(P)$ solutions.

As was mentioned above $\psi_{oT'}^m(P)$ is discontinuous at $\eta = 0$ unless $T' - 1 - M$ is a non-negative integer. Even though the other wave functions are discontinuous at $\eta = 0$, it is possible that there exists a linear combination of them which is continuous at $\eta = 0$.

We will be concerned with solutions odd in the variable η , of the form

$$\psi_0^m(s, \eta, \phi) = \sum_p \alpha_p \psi_{0p}^m(P), \quad (3.1)$$

where p increases by integral values only. The coefficients α_p functions of kd only, are normalized such that there exists at least one coefficient which does not vanish when k is zero. Let

$$[\psi_0^m(s, \eta, \phi)]_{k=0} = \sum_{p=T'}^{T'+N'} \alpha_p [\psi_{0p}^m(P)]_{k=0}, \quad (3.2)$$

where N' is a non-negative integer. We require $\psi_0^m(s, \eta, \phi)$ to be continuous at $\eta = 0$ for every s and kd in the ranges $1 \leq s \leq \infty$ and $0 \leq kd \leq \infty$. Hence a necessary condition for continuity is that the Laplace portion of $\psi_0^m(s, \eta, \phi)$ must be continuous at $\eta = 0$. Since $\psi_0^m(s, \eta, \phi)$ is an odd function of η , the following must hold for η approaching zero

$$[\psi_0^m(s, \eta, \phi)]_{k=0} \sim 0(\eta) \quad 1 \leq s \leq \infty. \quad (3.3)$$

Now it can be shown that $[\psi_{oT'}^m(P)]_{k=0}$ has the value as η approaches $+\infty$

$$[\psi_{oT'}^m(P)]_{k=0} \sim \left(\frac{s-1}{s+1} \right)^{M/2} \frac{(2\pi)^{1/2}(s-1)^{1/2-T'}}{\Gamma(-T'+1+M)\Gamma(T'+.5)} + 0(\eta). \quad (3.4)$$

The term independent of η in (3.4) is zero only if $T' - 1 - M$ or $-T' - \frac{1}{2}$ is a non-negative integer. Because of the factor $(s-1)^{-T'}$ there is no linear combination of $\psi_{oT'}^m(P)$ which will satisfy (3.3) unless $T' - 1 - M$ or $-T' - \frac{1}{2}$ is a non-negative integer. Hence in (3.2) the lower limits of summation may have the values $M+1, M+2, \dots$ or $-\frac{1}{2}, -\frac{3}{2}, \dots$. In the latter case the upper limit of summation is $-\frac{1}{2}$.

The even solutions of $\eta, \psi_{eT}^m(P)$ have all been shown to be uniformly convergent for the region $1 \leq s \leq \infty$ and $0 \leq \eta \leq 2\pi$, and hence are continuous everywhere in the region. However, if we differentiate term by term the series given by (2.24) with respect

to the variable η , the resulting series can be shown to be uniformly convergent in the region $1 \leq s \leq \infty$ and $0 < \eta < 2\pi$ and discontinuous at $\eta = 0$ or 2π , except when $T - M$ is a non-negative integer. In this case the differential series is continuous at $\eta = 0$ and $\eta = 2\pi$.

We will be concerned with obtaining a necessary condition for continuity of the partial derivative with respect to η for the even solutions $\psi_e^m(s, \eta, \phi)$ where

$$\psi_e^m(s, \eta, \phi) = \sum_p \beta_p \psi_{ep}^m(P) \quad (3.5)$$

and which has the value when k vanishes

$$[\psi_e^m(s, \eta, \phi)]_{k=0} = \sum_{p=T}^{T+N} \beta_p [\psi_{ep}^m(P)]_{k=0}. \quad (3.6)$$

Using the asymptotic relation for $\eta \rightarrow +0$

$$\left[\frac{\partial}{\partial \eta} \psi_{eT}^m(P) \right]_{k=0} \sim -\left(\frac{s-1}{s+1}\right)^{M/2} \frac{(2\pi)^{1/2}(s-1)^{-1/2-T}}{\Gamma(-T+M)\Gamma(T+.5)} + O(\eta) \quad (3.7)$$

we can deduce in a manner similar to the above that the lower limit of summation in (3.6) may be $M, M+1, M+2, \dots$ or $-\frac{1}{2}, -3/2, \dots$, and in the latter case the upper limit of summation is $-\frac{1}{2}$.

Imposing the above conditions of continuity on T and T' restricts the number of solutions.

To simplify further work $\psi_{eN}^m(P)$ and $\psi_{0N}^m(P)$ shall be defined by the relations

$$\psi_{eN}^m(P) = (kd)^N \psi_{e(M+N)/2}^m(P), \quad (3.8)$$

$$\psi_{0N}^m(P) = 2^{-1/2} (kd)^N \psi_{0(M+N+1)/2}^m(P), \quad (3.9)$$

where in (3.8) T is replaced by $(M+N)/2$ and in (3.9) T' is replaced by $(M+N+1)/2$.

Later on we will be concerned with wave functions of the form

$$\sum_{r=N}^{\infty} \alpha_r \psi_{0r}^*(P), \quad (3.10)$$

$$\sum_{r=N}^{\infty} \beta_r \psi_{er}^*(P), \quad (3.11)$$

where α_r and β_r are independent of kd . Substituting in the expressions (3.8) and (3.9) and normalizing we see that on employing the necessary conditions for continuity we obtain for (3.11)

$$N = M + 2l \quad \text{or} \quad -N - 1 = M + 2l$$

and for (3.10)

$$N = M + 2l + 1 \quad \text{or} \quad -N - 1 = M + 2l + 1,$$

where $l = 0, 1, 2, 3, \dots$.

4. Asymptotic value of $\psi_{eN}^m(P), \psi_{0N}^m(P)$ when $d \rightarrow 0$. The main problem is to find the linear combination of $\psi_{eN}^m(P)$ and $\psi_{0N}^m(P)$ which represents outgoing radiation from a ring source. A necessary condition for this is that the wave function represent radiation from a point source when the radius d of the ring approaches zero. Hence

one must first consider the asymptotic values

$$\lim_{d \rightarrow 0} \psi_{eN}^{m*}(R, \theta, \phi),$$

$$\lim_{d \rightarrow 0} \psi_{0N}^{m*}(R, \theta, \phi),$$

where (R, θ, ϕ) are spherical polar coordinates and before taking the limit, the toroidal coordinates (s, η, ϕ) are replaced by (R, θ, ϕ) .

Since

$$\rho = \frac{d(s^2 - 1)^{1/2}}{(s - \cos \eta)}, \quad z = \frac{d \sin \eta}{(s - \cos \eta)}, \quad (4.1)$$

where (ρ, z, ϕ) are cylindrical polar coordinates, one obtains

$$\frac{R^2}{d^2} = \frac{(\rho^2 + z^2)}{d^2} = \frac{(s + \cos \eta)}{(s - \cos \eta)} \quad (4.2)$$

and

$$\tan \theta = \frac{(s^2 - 1)^{1/2}}{\sin \eta}. \quad (4.3)$$

Eliminating η from (4.2) and (4.3), one can obtain an expression for s in terms of R and θ . Hence the following are obtained when d approaches zero

$$\left. \begin{aligned} (s^2 - 1)^{1/2} &\sim \frac{d}{R} 2 \sin \theta \\ s &\sim 1 + \frac{d^2}{R^2} 2 \sin^2 \theta \end{aligned} \right\}. \quad (4.4)$$

Similarly one obtains

$$\left. \begin{aligned} (s - \cos \eta) &\sim 2d^2 R^{-2} \\ \sin \eta &\sim 2dR^{-1} \cos \theta \end{aligned} \right\}. \quad (4.5)$$

The asymptotic values given by (4.4) and (4.5) hold not only when R is fixed and d approaches zero, but when d is fixed and R approaches infinity.

Using the above the following limits can be calculated

$$\left. \begin{aligned} \lim_{d \rightarrow 0} \psi_{eN}^{m*}(R, \theta, \phi) &= H_N^m(R, \theta, \phi) \\ \lim_{d \rightarrow 0} \psi_{0N}^{m*}(R, \theta, \phi) &= H_N^m(R, \theta, \phi) \end{aligned} \right\} 0 < \theta < \frac{\pi}{2}, \quad (4.6)$$

where

$$H_N^m(R, \theta, \phi) = 2^{(M-N)/2} e^{im\phi} \sum_{t=0}^{\infty} \frac{(-1)^t (kR)^{2t+N}}{2^t t!} \sin \theta' P_{N+t}^{-M-t} (\cos \theta). \quad (4.7)$$

For present purposes the limit is only required for the range $0 < \theta < \pi/2$.

5. Solutions of the wave equation representing radiations from a ring source. Having obtained the asymptotic values of $\psi_{eN}^{m*}(P)$ and $\psi_{0N}^{m*}(P)$ we now find the linear combination ψ_N^M representing a solution of the Helmholtz equation, satisfying the radia-

tion condition, and possessing a ring singularity (i.e. singular at the limiting surface $s = \infty$). This is done by using the necessary condition that ψ_N^M represent radiation from a point source when $d \rightarrow 0$. No further restriction is placed if it is required that

$$\lim_{d \rightarrow 0} \psi_N^M = e^{im\phi} h_N^{(1)}(kR) P_N^M(\cos \theta), \quad (5.1)$$

where N is an integer.

Since

$$h_N^{(1)}(kR) P_N^M(\cos \theta) = j_N(kR) P_N^M(\cos \theta) + i(-1)^{N+1} j_{-N-1}(kR) P_{-N-1}^M(\cos \theta)$$

it can be seen that the desired linear combination has the form

$$\psi_N^M = \Psi_N^M + i(-1)^{N+1} \Psi_{-N-1}^M$$

where

$$\Psi_N^M = \sum_r C_r(N, M) \Psi_r^m(P). \quad (5.2)$$

Thus it is seen that (5.1) can be replaced by

$$\lim_{d \rightarrow 0} \Psi_N^M = e^{im\phi} j_N(kR) P_N^M(\cos \theta). \quad (5.3)$$

Using (4.6) and (5.2) this reduces to

$$\sum_r C_r(N, M) H_r^m(R, \theta, \phi) = e^{im\phi} j_N(kR) P_N^M(\cos \theta) \quad 0 < \theta < \frac{\pi}{2}. \quad (5.4)$$

Equation (5.4) determines the unknown constants $C_r(N, M)$. It is an identity in the variables (R, θ, ϕ) . In solving for $C_r(N, M)$ we shall consider θ to be in the range $0 < \theta < \pi/2$. This will be sufficient for present purposes.

The right hand side of (5.4) is of the form

$$\frac{(\pi)^{1/2}}{2} \sum_{p=0}^{\infty} \frac{(-1)^p \left(\frac{kR}{2}\right)^{N+2p}}{(p)! \Gamma(p+N+3/2)} P_N^M(\cos \theta) e^{im\phi}. \quad (5.5)$$

This is a power series in (kR) with lowest power N and with all terms in the power series of the same parity as N . Thus considering the expression for H_r^m given by (4.7), it is obvious that the left hand side of (5.4) must be of the form

$$\sum_{r=0}^{\infty} C_{N+2r}(N, M) H_{N+2r}^m(R, \theta, \phi) \quad (5.6)$$

and on substitution of the expression given by (4.12) this becomes

$$e^{im\phi} 2^{(M-N)/2} (kR)^N \sum_{p=0}^{\infty} \frac{(kR)^{2p}}{2^p} \sum_{r=0}^p \frac{C_{N+2r}(N, M) \cdot (-1)^{p-r}}{\Gamma(p-r+1)} \sin \theta^{p-r} P_{N+p+r}^{-M-p+r}(\cos \theta). \quad (5.7)$$

Substitute the expression given by (5.5) and (5.7) in (5.4), divide out $(kR)^N e^{im\phi}$, and equate coefficients of $(kR)^{2p}$ to obtain

$$\begin{aligned} \frac{(\pi)^{1/2} P_N^M(\cos \theta)}{(p)! \Gamma(p+N+3/2) 2^{N+2p+1}} &= 2^{(M-N)/2-p} \sum_{r=0}^p \frac{C_{N+2r}(N, M) \cdot (-1)^r}{\Gamma(p-r+1)} \sin \theta^{p-r} \\ &\times P_{N+p+r}^{-M-p+r}(\cos \theta), \quad \text{where } p = 0, 1, 2, 3, \dots \end{aligned} \quad (5.8)$$

The constants $C_{N+2r}(N, M)$ are determined from the infinite set of equations (5.8). To solve for $C_{N+2r}(N, M)$ the following result is used

$$\frac{P_N^{-M}(\cos \theta)}{2^p(p)!\Gamma(p+N+3/2)} = \sum_{r=0}^p \frac{\sin \theta^{p-r} P_{N+p+r}^{-M-p+r}(\cos \theta)}{2^r(p-r)!(r)!\Gamma(N+3/2+r)}. \quad (5.9)$$

The proof of the above expression is found in [6].

Using the fact that

$$P_N^M(\cos \theta) = (-1)^M \frac{\Gamma(N+M+1)}{\Gamma(N-M+1)} P_N^{-M}(\cos \theta) \quad (5.10)$$

and the identity given by (5.9) one immediately obtains

$$C_{N+2r}(N, M) = (\pi)^{1/2} \frac{\Gamma(N+M+1)(-1)^{r+M}}{\Gamma(N-M+1)(r)!\Gamma(N+3/2+r) 2^{(M+N)/2+r+1}}. \quad (5.11)$$

So now the coefficients $C_r(N, M)$ in (5.2) have been found such that Eq. (5.3) holds. Since the limiting process of d approaching zero is independent of whether $\psi_{er}^{m*}(P)$ or $\psi_{0r}^{m*}(P)$ are used in (5.2), one has two separate solutions for Ψ_N^M , odd and even solutions in the variable η . Define the even solutions by S_N^M and the odd solutions by T_N^M as follows:

$$e^{im\phi} S_N^M = \sum_{r=0}^{\infty} C_{N+2r}(N, M) \psi_{eN+2r}^{m*}(P), \quad (5.12)$$

$$e^{im\phi} T_N^M = \sum_{r=0}^{\infty} C_{N+2r}(N, M) \psi_{0N+2r}^{m*}(P). \quad (5.13)$$

Applying the necessary condition for continuity of the function and its derivatives as given in Sec. (3), it is seen that in (5.12) the subscript N is such that

$$N = M + 2l \quad \text{or} \quad -N - 1 = M + 2l$$

and the subscript N in (5.13) is such that

$$N = M + 2l + 1 \quad \text{or} \quad -N - 1 = M + 2l + 1,$$

where $l = 0, 1, 2, \dots$

S_N^M and T_N^M have the following properties

$$\lim_{d \rightarrow 0} \begin{cases} S_N^M = j_N(kR) P_N^M(\cos \theta) \\ T_N^M = j_N(kR) P_N^M(\cos \theta) \end{cases} \left| \begin{array}{l} 0 < \theta < \frac{\pi}{2} \\ \end{array} \right.$$

Hence the functions V_N^M and W_N^M defined as follows,

$$V_N^M = S_N^M + i(-1)^{N+1} S_{-N-1}^M, \quad (5.14)$$

$$W_N^M = T_N^M + i(-1)^{N+1} T_{-N-1}^M, \quad (5.15)$$

have the property

$$\lim_{d \rightarrow 0} \begin{cases} V_N^M = h_N^{(1)}(kR) P_N^M(\cos \theta) \\ W_N^M = h_N^{(1)}(kR) P_N^M(\cos \theta) \end{cases} \left| \begin{array}{l} 0 < \theta < \frac{\pi}{2} \\ \end{array} \right.$$

Thus we have a set of functions $e^{im\phi} V_N^M(s, \eta)$ and $e^{im\phi} W_N^M(s, \eta)$ which satisfy the necessary condition for outgoing radiation. Their analytic properties of continuity convergence, singularities etc., must be investigated.

The first thing that is required is to evaluate (5.12) and (5.13). From (5.12) and (5.11) it is seen that

$$e^{im\phi} S_N^M = \frac{\Gamma(N + M + 1)(\pi)^{1/2}(-1)^m}{\Gamma(N - M + 1)2^{(M+N)/2+1}} \sum_{r=0}^{\infty} \frac{(-1)^r \psi_{eN+2r}^{m*}(P)}{2^r(r)\Gamma(N + 3/2 + r)}, \quad (5.16)$$

where by (3.8), $\psi_{eN+2r}^{m*}(P)$ is obtained from (2.24) by replacing T by $(N + M)/2 + r$ and multiplying the resulting series by $(kd)^{N+2r}$.

The expression given by (5.16) can be simplified and the following is obtained

$$S_N^M = \frac{\Gamma(N + M + 1)(\pi)^{1/2}(-1)^M(kd)^N}{\Gamma(N - M + 1)2^{(M+N)/2+1}(s - \cos \eta)^{(M+N)/2}} \sum_{p=0}^{\infty} \frac{(kd)^{2p}(-1)^p K_p^1}{(s - \cos \eta)^p 2^p}, \quad (5.17)$$

where

$$K_p^1 = \frac{1}{(p)!\Gamma(N + 3/2 + p)} \sum_{r=0}^{\infty} \frac{\left(\frac{M-N}{2}\right)_r \left(\frac{N+M+1}{2}\right)_r (s^2 - 1)^{r/2}}{(r)!(s - \cos \eta)^r} \cdot P_{p+(N+M)/2}^{-M-r}(s). \quad (5.18)$$

Since N is specified for S_N^M such that $N = M + 2l$ or $-M - 2l - 1$, and l is a positive integer or zero, the series in (5.18) terminates when $r = l$.

Similarly one can reduce the expression for T_N^M to the following

$$T_N^M = \frac{\Gamma(N + M + 1)(\pi)^{1/2}(-1)^M \sin \eta (kd)^N}{\Gamma(N - M + 1)2^{(M+N+3)/2}(s - \cos \eta)^{(M+N+1)/2}} \sum_{p=0}^{\infty} \frac{(kd)^{2p}(-1)^p K_p^2}{2^p(s - \cos \eta)^p}, \quad (5.19)$$

where

$$K_p^2 = \frac{1}{(p)!\Gamma(p + 3/2 + N)} \sum_{r=0}^{\infty} \frac{\left(\frac{M-N+1}{2}\right)_r \left(\frac{M+N}{2} + 1\right)_r (s^2 - 1)^{r/2}}{(r)!(s - \cos \eta)^r} \cdot P_{p+(M+N-1)/2}^{-M-r}(s). \quad (5.20)$$

Since N is specified for T_N^M to have the values $N = M + 2l + 1$ or $-M - 2l - 2$, the series in (5.20) terminates when $r = l$.

Hence we see that S_N^M and T_N^M are comprised of two series one finite and the other infinite. Hence for convergence, we only need to consider summation over p . Using the inequality (2.28) an absolutely convergent dominant series can be found for the series

$$\sum_p^{\infty} \frac{(kd)^{2p} P_{p+(N+M)/2}^{-M-r}(s)}{2^p(s - \cos \eta)^p (p)!\Gamma(N + 3/2 + p)}.$$

Hence it can be shown that the series expression in (5.17) is absolutely and uniformly convergent for the region $1 \leq s \leq \infty$, $0 \leq \eta \leq 2\pi$ and $0 \leq |kd| \leq \infty$. If the series (5.17) is differentiated term by term by either s or η , the resulting series is also uniformly convergent for the same region.

The same remarks hold true for the series expansion (5.19) given for T_N^M .

6. Integral representation of $V_{M+2l}^M(s, \eta)$ and $W_{M+2l+1}^M(s, \eta)$. Before the asymptotic values for large R can be obtained, we must obtain an integral representation for each of the functions.

Using the following integral expression for the associated Legendre function $P_n^\mu(s)$ where $\mu < \frac{1}{2}$

$$P_n^\mu(s) = \frac{2^\mu (s^2 - 1)^{-\mu/2}}{(\pi)^{1/2} \Gamma(\frac{1}{2} - \mu)} \int_0^\pi [s + (s^2 - 1)^{1/2} \cos t]^{n+\mu} (\sin t)^{-2\mu} dt \quad (6.1)$$

we obtain

$$\frac{(s^2 - 1)^{r/2} P_n^{-M-r}(s)}{(s - \cos \eta)^r} = \frac{2^{-M} (s^2 - 1)^{M/2}}{(\pi)^{1/2} \Gamma(\frac{1}{2} + M + r)} \int_0^\pi \frac{Z^r (\sin t)^{2M} dt}{[s + (s^2 - 1)^{1/2} \cos t]^{M+r}}, \quad (6.2)$$

where

$$Z = \frac{(s^2 - 1) (\sin t)^2}{2(s - \cos \eta)[s + (s^2 - 1)^{1/2} \cos t]}. \quad (6.3)$$

From (5.18) and (6.2) we thus obtain

$$K_p^1 = \frac{2^{-M} (s^2 - 1)^{M/2} (\pi)^{-1/2}}{(p)! \Gamma(N + 3/2 + p) \Gamma(\frac{1}{2} + M)} \cdot \int_0^\pi \frac{{}_2F_1\left(\frac{M-N}{2}, \frac{N+M+1}{2}; \frac{1}{2} + M; Z\right) (\sin t)^{2M} dt}{[s + (s^2 - 1)^{1/2} \cos t]^{(M-N)/2-p}}. \quad (6.4)$$

The range of Z is $0 \leq Z \leq 1$, Z being unity when $\cos \eta = 1$ and $s + (s^2 - 1)^{1/2} \cos t = 1$. Since we are considering the case where $N = M + 2l$ or $-N - 1 = M + 2l$ the hypergeometric function in the expression (6.4) is a polynomial with finite argument. Thus the interchange of order of summation and integration is valid. Now in the expression for S_N^M (5.17) substitute expression (6.4) for K_p^1 . Interchange the order of summation and integration. Noting that

$$j_N(X) = \sum_{p=0}^{\infty} \frac{(-1)^p \left(\frac{x}{2}\right)^{2p+N}}{(p)! \Gamma(N + 3/2 + p)} \frac{(\pi)^{1/2}}{2} \quad (6.5)$$

$$= \frac{(\pi)^{1/2} (k d)^N 2^{-N/2-1}}{(s - \cos \eta)^{N/2}} \sum_{p=0}^{\infty} \frac{(-1)^p (k d)^{2p} [s + (s^2 - 1)^{1/2} \cos t]^{p+N/2}}{2^p (s - \cos \eta)^p (p)! \Gamma(N + 3/2 + p)}, \quad (6.6)$$

where

$$X = kd 2^{1/2} \left[\frac{s + (s^2 - 1)^{1/2} \cos t}{s - \cos \eta} \right]^{1/2} \quad (6.7)$$

it is seen that the infinite series in the integrand is uniformly convergent for $\infty \geq s \geq 1$ and $0 \leq t \leq \pi$. Hence the following holds

$$S_N^M = \frac{\Gamma(N + M + 1) (-2)^M (\pi)^{-1/2}}{\Gamma(N - M + 1) \Gamma(\frac{1}{2} + M)} \cdot \int_0^\pi {}_2F_1\left(\frac{M-N}{2}, \frac{N+M+1}{2}; \frac{1}{2} + M; Z\right) \cdot j_N(X) Z^{M/2} (\sin t)^M dt \quad 1 \leq s < \infty. \quad (6.8)$$

When $N = M + 2l$ the integral in (6.8) will hold also for $s = \infty$. From (5.14) we have

$$V_{M+2l}^M = \frac{(2M+2l)!(-2)^{-M}(\pi)^{-1/2}}{(2l)!\Gamma(\frac{1}{2}+M)} \int_0^\pi {}_2F_1(-l, M+l+\frac{1}{2}; \frac{1}{2}+M; Z) h_{M+2l}^{(1)}(X) \times Z^{M/2} (\sin t)^M dt, \quad (6.9)$$

where $h_N^{(1)}(X)$ is the spherical Hankel function of the first kind. In a similar manner the following integral expression may be obtained for T_N^M where $N = M + 2l + 1$ or $-N - 1 = M + 2l + 1$,

$$T_N^M = \frac{\Gamma(N+M+1)(-2)^{-M} \sin \eta(s^2 - 1)^{-1/2}}{\Gamma(N-M+1)(\pi)^{1/2}\Gamma(M+\frac{1}{2})} \cdot \int_0^\pi {}_2F_1\left(\frac{M-N+1}{2}, \frac{M+N}{2}+1; \frac{1}{2}+M; Z\right) \cdot j_N(X) Z^{(M+1)/2} (\sin t)^{M-1} dt. \quad (6.10)$$

The integral expression in (6.10) holds in the ranges $1 \leq s \leq \infty$ for $N = M + 2l + 1$ and $1 \leq s < \infty$ for $N = -M - 2l - 2$. We also obtain the integral expression for W_{M+2l+1}^M which holds for $1 \leq s < \infty$

$$W_{M+2l+1}^M = \frac{(2M+2l+1)!(-2)^{-M} \sin \eta(s^2 - 1)^{-1/2}}{(2l+1)!(\pi)^{1/2}\Gamma(\frac{1}{2}+M)} \cdot \int_0^\pi {}_2F_1(-l, M+l+3/2; \frac{1}{2}+M; Z) h_{M+2l+1}^{(1)}(X) Z^{(M+1)/2} (\sin t)^{M-1} dt. \quad (6.11)$$

7. Asymptotic values of $V_{M+2l}^M(s, \eta)$ and $W_{M+2l+1}^M(s, \eta)$ for $R \rightarrow \infty$. Having obtained the integral formulation of the wave functions, we are now in a position to investigate their asymptotic values when R approaches ∞ .

From (4.4) and (4.5) the following asymptotic values are obtained for R approaching ∞

$$\frac{(s^2 - 1)^{1/2}}{s} \sim \frac{d}{R} \left[2 \sin \theta + O\left(\frac{1}{R}\right) \right], \quad (7.1)$$

$$\frac{s - \cos \eta}{s} \sim \frac{2d^2}{R^2} \left[1 + O\left(\frac{1}{R^2}\right) \right]. \quad (7.2)$$

Hence the asymptotic values of Z and X for large R are

$$Z \sim \sin^2 t \sin^2 \theta + O\left(\frac{1}{R}\right) \quad (7.3)$$

and

$$X \sim kR + kd \cos t \sin \theta + O\left(\frac{1}{R}\right). \quad (7.4)$$

Insert the values (7.3) and (7.4) into (6.9) and using the fact that when R approaches ∞

$$h_{M+2l}^{(1)}(X) \sim (-i)^{M+2l+1} \frac{\exp(ikR + ikd \cos t \sin \theta)}{kR} \quad (7.5)$$

we have

$$V_{M+2l}^M \sim (-i)^{M+2l+1} \frac{e^{ikR}}{kR} R_{M+2l}^M (\cos \theta), \quad (7.6)$$

where

$$R_{M+2l}^M(\cos \theta) = \frac{(2M+2l)!(-1)^M(\pi)^{-1/2}}{(2l)!(2)^M \Gamma(M+.5)} \times \int_0^\pi {}_2F_1(-l, M+l+\frac{1}{2}; \frac{1}{2}+M; \sin t^2 \sin \theta^2) \sin \theta^M \sin t^{2M} e^{ikd \cos t \sin \theta} dt. \quad (7.7)$$

Expand the hypergeometric series in (7.7) and use the relation

$$\Gamma(n + \frac{1}{2}) J_n(z) = \pi^{-1/2} \left(\frac{z}{2}\right)^n \int_0^\pi e^{iz \cos t} \sin t^{2n} dt \quad (7.8)$$

to obtain

$$R_{M+2l}^M(\cos \theta) = \frac{(2M+2l)!}{(2l)!(-kd)^M} \sum_{r=0}^l \frac{(-l)_r (M+l+\frac{1}{2})_r}{(r)!} \left(\frac{2 \sin \theta}{kd}\right)^r J_{M+r}(kd \sin \theta). \quad (7.9)$$

To obtain the asymptotic expansion for $W_{M+2l+1}^M(s, \eta)$ the following result is needed [from (4.3)]

$$\sin \eta(s^2 - 1)^{-1/2} = \cot \theta. \quad (7.10)$$

In a manner similar to the above, the asymptotic value of $W_{M+2l+1}^M(s, \eta)$ can be obtained for R approaching ∞

$$W_{M+2l+1}^M(s, \eta) \sim (-i)^{M+2l+2} \frac{e^{ikR}}{kR} R_{M+2l+1}^M(\cos \theta), \quad (7.11)$$

where

$$R_{M+2l+1}^M(\cos \theta) = \frac{(2M+2l+1)! \cos \theta}{(2l+1)!(-kd)^M} \sum_{r=0}^l \frac{(-l)_r (M+l+3/2)_r}{(r)!} \left(\frac{2 \sin \theta}{kd}\right)^r J_{M+r}(kd \sin \theta). \quad (7.12)$$

From (7.6) and (7.11) it is seen that the wave functions $e^{im\phi} V_{M+2l}^M(s, \eta)$ and $e^{im\phi} W_{M+2l+1}^M(s, \eta)$ will satisfy the radiation condition. We shall consider the case where d approaches zero. It can be shown that the asymptotic values for Z and X given by (7.3) and (7.4) will hold when d vanishes. Hence we have

$$\lim_{d \rightarrow 0} V_{M+2l}^M = h_{M+2l}^{(1)}(kR) \lim_{d \rightarrow 0} R_{M+2l}^M(\cos \theta). \quad (7.13)$$

But

$$\begin{aligned} \lim_{d \rightarrow 0} R_{M+2l}^M(\cos \theta) &= \frac{(2M+2l)!(-1)^M}{(2l)!(M)!} \cdot \left(\frac{\sin \theta}{2}\right)^M {}_2F_1(-l, M+l+\frac{1}{2}; M+1; \sin^2 \theta) \\ &= \frac{(2M+2l)!(-1)^M}{(2l)!} P_{M+2l}^{-M}(\cos \theta) \end{aligned} \quad (7.14)$$

$$= P_{M+2l}^M(\cos \theta) \quad 0 \leq \theta \leq \pi. \quad (7.15)$$

We see that $R_{M+2l}^M(\cos \theta)$ is identical with the associated Legendre function $P_{M+2l}^M(\cos \theta)$ when d vanishes. In a similar manner it can be shown that $R_{M+2l+1}^M(\cos \theta)$ becomes identical to $P_{M+2l+1}^M(\cos \theta)$ when d is zero. Hence the toroidal wave functions $e^{im\phi} V_{M+2l}^M(s, \eta)$ and $e^{im\phi} W_{M+2l+1}^M(s, \eta)$ are identical to the spherical polar wave functions

when d vanishes. Thus when d vanishes they represent radiations from a point source.

The remaining detail to be considered is the nature of the singularities of the wave function at $s = \infty$.

8. Asymptotic behaviour of $V_{M+2l}^M(s, \eta)$, $W_{M+2l+1}^M(s, \eta)$ as $s \rightarrow \infty$. In order to investigate the asymptotic behaviour of $V_{M+2l}^M(s, \eta)$ and $W_{M+2l+1}^M(s, \eta)$ we must investigate the functions S_{M+2l}^M , $S_{-M-2l-1}^M$, T_{M+2l+1}^M and $T_{-M-2l-2}^M$ separately.

For S_{M+2l}^M and T_{M+2l+1}^M use the integral expressions. We have shown that the expressions hold for $1 \leq s \leq \infty$ and since the integrand is finite for $s = \infty$ we may use the integral expression to determine the asymptotic values.

From (6.3) and (6.7) the asymptotic values of Z and X become as s approaches ∞

$$Z \sim 2^{-1}(1 - \cos t), \quad (8.1)$$

$$X \sim kd 2^{1/2}(1 + \cos t)^{1/2}. \quad (8.2)$$

Hence we obtain the following for s approaching ∞

$$S_{M+2l}^M \sim \text{constant}, \quad (8.3)$$

$$T_{M+2l+1}^M \sim \frac{\sin \eta}{s} \text{constant}. \quad (8.4)$$

It is more difficult to find the asymptotic values for $S_{-M-2l-1}^M$ and $T_{-M-2l-2}^M$. For simplification of work, the functions $F_p(s, \eta; l, M)$ and $G_p(s, \eta; l, M)$ are defined as follows

$$\left. \begin{aligned} F_p(s, \eta; l, M) &= (s - \cos \eta)^{l+1/2} \sum_{r=0}^l \frac{(-l)_r (M + l + \frac{1}{2})_r}{(r)! (s - \cos \eta)^r} (s^2 - 1)^{r/2} P_{p-l-1/2}^{-M-r}(s) \\ G_p(s, \eta; l, M) &= \sin \eta (s - \cos \eta)^{l+1/2} \sum_{r=0}^l \frac{(-l)_r (M + l + 3/2)_r}{(r)! (s - \cos \eta)^r} \\ &\quad \cdot (s^2 - 1)^{r/2} P_{p-l-3/2}^{-M-r}(s) \end{aligned} \right\}. \quad (8.5)$$

Thus from (5.17), (5.18) and (8.5) it is seen that

$$\begin{aligned} S_{-M-2l-1}^M(s, \eta) &= \frac{\Gamma(2M + 2l + 1)(\pi)^{1/2}(kd)^{-M-2l-1}}{\Gamma(2l + 1)2^{-l+1/2}(-1)^M} \\ &\quad \cdot \sum_{p=0}^{\infty} \frac{(-1)^p (kd)^{2p} F_p(s, \eta; l, M)}{(p)! 2^p \Gamma(-M - 2l + p + \frac{1}{2})(s - \cos \eta)^p} \end{aligned} \quad (8.6)$$

and from (5.19) and (5.20)

$$\begin{aligned} T_{-M-2l-2}^M(s, \eta) &= \frac{\Gamma(2M + 2l + 2)(\pi)^{1/2}(kd)^{-M-2l-2}}{\Gamma(2l + 2)2^{-l+1/2}(-1)^M} \\ &\quad \cdot \sum_{p=0}^{\infty} \frac{(-1)^p (kd)^{2p} G_p(s, \eta; l, M)}{(p)! 2^p \Gamma(-M - 2l + p - \frac{1}{2})(s - \cos \eta)^p}. \end{aligned} \quad (8.7)$$

Now multiply $S_{-M-2l-1}^M(s, \eta)$ by $(kd)^{M+2l+1} e^{im\phi}$ and let k approach zero. The only term which is non-vanishing is the term involving $F_0(s, \eta; l, M)$. Thus we have the following

$$\begin{aligned} \{(kd)^{M+2l+1} e^{im\phi} S_{-M-2l-1}^M\}_{k=0} &= \frac{\Gamma(2M + 2l + 1)(\pi)^{1/2} 2^{l-1/2} (-1)^M e^{im\phi}}{\Gamma(2l + 1) \Gamma(-M - 2l + \frac{1}{2})} F_0(s, \eta; l, M). \end{aligned} \quad (8.8)$$

But any solution of the wave equation when $k = 0$ is a solution of the Laplace equation. Thus $e^{im\phi} F_0(s, \eta; l, M)$ is a solution of the Laplace equation. Now $F_0(s, \eta; l, M)$ is a power series of $(s - \cos \eta)$ up to the l th power, multiplied by $(s - \cos \eta)^{1/2}$ and is non-singular when $s = 1$. Since the complete set of solutions of the Laplace equation which are non-singular at $s = 1$ are

$$\text{and } \left. \begin{aligned} & e^{im\phi} (s - \cos \eta)^{1/2} P_{n-1/2}^{-M}(s) \cos n\eta \\ & e^{im\phi} (s - \cos \eta)^{1/2} P_{n-1/2}^{-M}(s) \sin n\eta \end{aligned} \right\} \quad (8.9)$$

then

$$F_0(s, \eta; l, M) = \sum_{r=0}^l a_r (s - \cos \eta)^{1/2} P_{r-1/2}^{-M}(s) \cos r\eta. \quad (8.10)$$

By a similar analysis it is seen that

$$G_0(s, \eta; l, M) = \sum_{r=1}^{l+1} b_r (s - \cos \eta)^{1/2} P_{r-1/2}^{-M}(s) \sin r\eta. \quad (8.11)$$

So the problem is to calculate a_r and b_r . To find the coefficients a_r multiply both sides of Eq. (8.10) by $(s^2 - 1)^{-M/2}$ and let s approach 1, using the fact that when s approaches 1

$$(s^2 - 1)^{(r-M)/2} P_n^{-M-r}(s) \sim \frac{2^{-M}(s-1)^r}{\Gamma(1+M+r)},$$

we obtain the equation

$$\sum_{r=0}^l a_r \cos r\eta = (1 - \cos \eta)^l. \quad (8.12)$$

But on the expansion of $(1 - \cos \eta)^r$ in a Fourier series we see that

$$a_r = \frac{\epsilon_r(2l)!(-1)^r}{(l-r)!(l+r)!2^{l-1}}, \quad (8.13)$$

where

$$\left. \begin{aligned} \epsilon_r &= 1 & \text{when } r = 1, 2, 3, \dots \\ \epsilon_r &= \frac{1}{2} & \text{when } r = 0. \end{aligned} \right\} \quad (8.14)$$

Now in a similar manner Eq. (8.11) may be reduced to the following

$$\sin \eta (1 - \cos \eta)^l = \sum_{r=1}^{l+1} b_r \sin r\eta. \quad (8.15)$$

Thus the coefficients b_r are derived from a Fourier analysis and b_r are

$$b_r = \frac{r(2l+1)l!(-1)^{r+1}}{(l+1+r)!(l+1-r)!2^{l-1}}. \quad (8.16)$$

Thus we obtain the following expressions for $F_0(s, \eta; l, M)$ and $G_0(s, \eta; l, M)$

$$F_0(s, \eta; l, M) = (s - \cos \eta)^{1/2} \sum_{r=0}^l \frac{\epsilon_r(2l)!(-1)^r}{(l-r)!(l+r)!2^{l-1}} P_{r-1/2}^{-M}(s) \cos r\eta, \quad (8.17)$$

$$G_0(s, \eta; l, M) = (s - \cos \eta)^{1/2} \sum_{r=1}^{l+1} \frac{r(2l+1)l!(-1)^{r+1}}{(l+1+r)!(l+1-r)!2^{l-1}} P_{r-1/2}^{-M}(s) \sin r\eta. \quad (8.18)$$

Using the following asymptotic values for the associated Legendre function

$$\left. \begin{aligned} P_{-1/2}^{-M}(s) &\sim \left(\frac{2}{\pi}\right)^{1/2} \frac{s^{-1/2}}{\Gamma(M + \frac{1}{2})} \{ \log_e(2s) - \gamma - \psi(M + \frac{1}{2}) \} \\ \text{where } \gamma &= 0.5772156649 \dots \quad \text{and} \quad \psi(z) = \frac{d \log_e \Gamma(z)}{dz} \\ P_n^{-M}(s) &\sim \frac{2^n \Gamma(n + \frac{1}{2}) s^n}{(\pi)^{1/2} \Gamma(1 + n + M)} \quad n > -\frac{1}{2} \end{aligned} \right\} \quad (8.19)$$

we see that the values of $F_0(s, \eta; l, M)$ and $G_0(s, \eta; l, M)$ become the following when s approaches ∞

$$F_0(s, \eta; l, M) \sim \frac{\epsilon_l(-1)^l}{2^{l-1}} s^{1/2} P_{l-1/2}^{-M}(s) \cos l\eta, \quad (8.20)$$

$$G_0(s, \eta; l, M) \sim \frac{(-1)^l s^{1/2}}{2^l} P_{l+1/2}^{-M}(s) \sin(l+1)\eta. \quad (8.21)$$

It can be shown [6] that

$$(s - \cos \eta)^{-p} F_p(s, \eta; l, M) \leq 0 \quad p = 1, 2, 3, \dots$$

and hence the dominant term in expression (8.6) given for $S_{-M-2l-1}^M$ is $F_0(s, \eta; l, M)$ for $s \gg 1$ provided that $kd < s$. Similarly it can be shown that $G_0(s, \eta; l, M)$ is the dominant term in expression (8.7) given for $T_{-M-2l-1}^M$.

Thus from (8.6) and (8.20) we have when s approaches ∞

$$S_{-M-2l-1}^M(s, \eta) \sim \frac{(2M + 2l)!(2\pi)^{1/2}(-1)^{M+l}\epsilon_l}{(2l)!\Gamma(-M - 2l + \frac{1}{2})(kd)^{M+2l+1}} s^{1/2} P_{l-1/2}^{-M}(s) \cos l\eta \quad (8.22)$$

and from (5.14), (8.22) and (8.3) we have

$$V_{M+2l}^M(s, \eta) \sim \frac{i(-1)^{l+1}(2\pi)^{1/2}(2M + 2l)!\epsilon_l}{(kd)^{M+2l+1}(2l)!\Gamma(-M - 2l + \frac{1}{2})} s^{1/2} P_{l-1/2}^{-M}(s) \cos l\eta. \quad (8.23)$$

From (8.7) and (8.21) we have

$$T_{-M-2l-2}^M(s, \eta) \sim \frac{(2M + 2l + 1)!(-1)^{M+l}(\pi/2)^{1/2}s^{1/2}}{(2l + 1)!\Gamma(-M - 2l - \frac{1}{2})(kd)^{M+2l+2}} P_{l+1/2}^{-M}(s) \sin(l+1)\eta \quad (8.24)$$

and hence

$$W_{M+2l+1}^M(s, \eta) \sim \frac{i(-1)^l(2M + 2l + 1)!(\pi/2)^{1/2}s^{1/2}}{(2l + 1)!\Gamma(-M - 2l - \frac{1}{2})(kd)^{M+2l+2}} P_{l+1/2}^{-M}(s) \sin(l+1)\eta. \quad (8.25)$$

9. Orthogonality and general discussion. We have obtained a set of solutions of the Helmholtz equation in toroidal coordinates which satisfy the radiation condition, are continuous and convergent in all space and possess a ring singularity. When the radius d of the ring described by the coordinate $s = \infty$ approaches zero, the toroidal wave functions become identical with spherical polar wave functions. Hence the toroidal wave functions $e^{im\phi} V_{M+2l}^M(s, \eta)$ and $e^{im\phi} W_{M+2l+1}^M(s, \eta)$ form a complete set. That is, any solution which is continuous and single-valued outside the torus $s = s_0$ (i.e. region $1 \leq s \leq s_0$) and satisfies the radiation condition and arbitrary boundary conditions on the torus can be represented by a linear combination of the above toroidal wave functions.

Now the question of orthogonality arises. Are the functions $e^{im\phi} V_{M+2l}^M(s, \eta)$ and $e^{im\phi} W_{M+2l+1}^M(s, \eta)$, orthogonal over the surface of every torus?

If we choose a weight function $p(s, \eta)$ which is an even function of the variable η , then we obtain the following

$$\int_0^{2\pi} \int_0^{2\pi} [e^{im\phi} V_{M+2l}^M(s, \eta)][e^{-im'\phi} V_{M'+2l'}^M(s, \eta)]p(s, \eta) d\eta d\phi = 0, \quad M \neq M' \quad (9.1)$$

$$\int_0^{2\pi} \int_0^{2\pi} [e^{im\phi} W_{M+2l+1}^M(s, \eta)][e^{-im'\phi} W_{M'+2l'+1}^M(s, \eta)]p(s, \eta) d\eta d\phi = 0, \quad M' \neq M \quad (9.2)$$

$$\int_0^{2\pi} \int_0^{2\pi} [e^{im\phi} V_{M+2l}^M(s, \eta)][e^{-im'\phi} W_{M'+2l'+1}^M(s, \eta)]p(s, \eta) d\eta d\phi = 0, \quad (9.3)$$

which holds for every torus $s = \text{constant}$. What can be said about the integrals?

$$\int_0^{2\pi} \int_0^{2\pi} [e^{im\phi} V_{M+2l}^M(s, \eta)][e^{-im\phi} V_{M+2l'}^M(s, \eta)]p(s, \eta) d\eta d\phi, \quad l \neq l' \quad (9.4)$$

$$\int_0^{2\pi} \int_0^{2\pi} [e^{im\phi} W_{M+2l+1}^M(s, \eta)][e^{-im\phi} W_{M+2l'+1}^M(s, \eta)]p(s, \eta) d\eta d\phi, \quad l \neq l'. \quad (9.5)$$

It can be shown that if $p = h_\phi$, then the integrals given by (9.4) and (9.5) do not vanish. In fact for $p = h_\phi$, it can be shown that there is no set of linear combinations of the wave functions $e^{im\phi} V_{M+2l}^M$ and $e^{im\phi} W_{M+2l+1}^M$ which form a complete orthogonal set over the surface of every torus. For $p \neq h_\phi$ nothing at present can be said about the vanishing of the integrals given by (9.4) and (9.5).

As it stands now one can say that, the wave functions $e^{im\phi} V_{M+2l}^M(s, \eta)$ and $e^{im\phi} W_{M+2l+1}^M(s, \eta)$ form a partially orthogonal set over every torus. To complete the orthogonal set, one must at present use the Hilbert-Schmidt process for every torus $s = s_0$. That is, for each torus $s = s_0$ and each value of m one must form an orthogonal set from the functions $e^{im\phi} V_M^M(s_0, \eta)$, $e^{im\phi} V_{M+2}^M(s_0, \eta)$, $e^{im\phi} V_{M+4}^M(s_0, \eta)$, ... using the Hilbert-Schmidt process, and also form an orthogonal set from the functions $e^{im\phi} W_{M+1}^M(s_0, \eta)$, $e^{im\phi} W_{M+3}^M(s_0, \eta)$, However the functions do form a complete orthogonal set over the surface of the limiting torus $s = s_0$ where $s_0 \gg 1$.

Apart from the difficulty involved in the problem of incomplete orthogonality which is a property of non-separability, the toroidal wave functions derived in this paper can be used to solve any problem of diffraction of acoustic waves by a torus or of a radiating torus. In fact they are useful in solving the vector wave equation, and already have been of practical value in electromagnetic problems.

Acknowledgement. The author gratefully acknowledges the suggestions and criticism of Dr. A. J. Coleman. The work was supported by the National Research Council of Canada through post-graduate scholarships.

REFERENCES

1. A. Erdelyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher transcendental functions*, vol. I, New York, 1953.
2. E. W. Hobson, *The theory of spherical and ellipsoidal harmonics*, Cambridge, 1931
3. A. Erdelyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher transcendental functions*, vol. II, New York, 1954.

4. J. Stratton, *Electromagnetic theory*, McGraw-Hill, 1941.
5. W. Magnus and F. Oberhettinger, *Functions of mathematical Physics*, (English ed. 1949)
6. V. H. Weston, *Solutions of the toroidal wave equation and their applications*, Ph.D. thesis, University of Toronto, 1956
7. V. H. Weston, *Solutions of the Helmholtz equation for a class of non-separable cylindrical and Rotational coordinate systems*, Quart. Appl. Math. **15**, 420 (1957)

BOOK REVIEWS*(Continued from p. 236)*

Neutron transport theory. By B. Davison. With the collaboration of J. B. Sykes. Oxford University Press, New York, 1957. xx 450 pp. \$12.00.

This book has for its central theme the study of the Boltzmann equation for neutron transport, the fundamental equation that describes the migration of neutrons through material media. It is evident that this study is of principal importance in the design of nuclear reactors. It is also of direct interest in astrophysics in the consideration of problems of radiative transfer which are very similar to those of neutron transport theory.

In this monograph the author has set for himself the task of giving a thorough review of all the important mathematical methods used in neutron transport theory. The subject matter is developed from first principles, and so a previous familiarity with the theory is not strictly necessary. Nevertheless, this book is not meant to serve as a first introduction to the subject. It is an advanced work which assumes a fair degree of mathematical maturity on the part of the reader; it requires comparatively little, however, in the way of a knowledge of nuclear physics and quantum mechanics.

The material covered in this work is conveniently divided into four parts. The first part is devoted to a precise formulation of the laws of neutron transport, the derivation of the basic integral and integro-differential equations, and the relation of stationary and time-dependent problems. The second part, by far the largest in the book, is concerned with the methods of solution of the transport equation in the constant cross-section approximation, i.e., for the case of one-group theory. Apart from a consideration of the few situations which can be treated exactly, this section develops in considerable detail the various approximation methods that have been proposed to date; the spherical harmonics method, in particular, is discussed at great length. The remainder of the book is concerned with energy-dependent problems for the cases of spectrum-regenerating media (Part III) and slowing-down media (Part IV).

This book fills an important gap in the literature, tying together, as it does, a considerable amount of material which had previously been available only in the form of technical reports and periodical articles. It is a serious mathematical work which is very well organized and very well written. It should prove extremely satisfying to theoretical physicists and applied mathematicians who are interested in an up-to-date account of the problems of neutron transport which goes far beyond the simplified discussion of neutron diffusion theory as it is usually presented in textbooks on nuclear reactors.

While one may question the nature of certain limitations on the scope of the subject which were imposed by the author (thus, the discussion is limited throughout to homogeneous media, with little or no reference being made to experimental methods or results), there is no doubt but that these have made possible a very well-knit and unified presentation; in any case, the book is already of considerable length. However, it would have been useful, especially in a comprehensive treatise such as this, if the author had documented this work much more extensively. The book is highly recommended.

DAVID FELDMAN

Elementary theory of angular momentum. By M. E. Rose. John Wiley & Sons, Inc., New York, 1957. 248 pp. \$10.00.

The quantum theory of angular momentum occupies a central position in present-day atomic and nuclear physics. Evidently, in the description of complicated systems, it is important to be able to separate out those aspects which can be traced directly to the existence in nature of certain fundamental symmetries from those which depend upon the detailed characteristics of the systems themselves (the specific shape of the nuclear force, for example). Notwithstanding the recent experimental verification that the so-called weak interactions are not invariant under spatial reflections, it still appears that rotational invariance is an absolute symmetry principle; hence, angular momentum is always conserved.

The book under review is an outgrowth of a series of lectures given recently by the author at the Oak Ridge National Laboratory. As is perhaps evident from the title, it is a textbook and not a treatise.

(Continued on p. 306)

VARIATIONAL PRINCIPLES FOR GUIDED ELECTROMAGNETIC WAVES IN ANISOTROPIC MATERIALS

BY

WALTER HAUSER

Lincoln Laboratory, M.I.T., Lexington, Mass.

Abstract. It is the purpose of this paper to develop a general method for obtaining approximate solutions to the problems of the propagation of waves in a guide which may only be partially filled with material having tensor electromagnetic properties. With the introduction of an appropriate dyadic Green's function we are able to obtain a formal solution to the problem in terms of an integral involving the field vectors in the perturbing rod. The reformulation of the original problem in terms of an integral equation enables us to construct a variational principle whose extremal value is insensitive to a great range of trial functions. The extremal value of the variational expression from which the integral equations for the field are derivable is shown to be proportional to $(\gamma_0^2 - \gamma^2)$, the difference between the square of the propagation constant of the wave in the empty and loaded guide. Normal modes in terms of which an expansion of the dyadic Green's function is obtained are defined. In the last section we demonstrate the ability of the variational method to improve the results obtained by perturbation methods obtaining a first order approximation for the propagation constant of a wave in a rectangular guide containing an infinite ferrite slab.

Introduction. With the use of ferrites in the design of microwave gyrators, circulators and nonreciprocal transmission systems, one has become interested in the problem of cavities and waveguides containing materials having tensor electromagnetic properties. In general, such a problem is extremely difficult and tedious. One is almost impelled to look for approximate solutions. In those problems where the material fills but a small part of the cavity or guide, or where the electromagnetic properties of the perturbing material differ but slightly from those of the empty guide or cavity, perturbation theory has been somewhat successful in obtaining first order approximations [1]. There exist a number of problems, however, to which perturbation theory yields poor or no results.

This paper is the first of a group of papers in which we shall concern ourselves with the development and application of a general method for obtaining approximate solutions to the problems of waveguides and cavities containing materials with tensor electromagnetic properties. The method is an extension of the work of Schwinger [2] on the problem of isotropic obstacles in cavities. With the introduction of an appropriate dyadic Green's function we are able to obtain a formal solution to the problem in terms of integrals involving the field vectors in the perturbing material. While the resulting integral equations are not any easier to solve, there exists the advantage of being able to construct stationary expressions for the quantities of interest from which the integral equations for the fields within the material are derivable. Consequently, we have a very powerful method for obtaining approximate solutions for these quantities. In this paper, for example, we concern ourselves with the problem of the propagation of waves in a guide which may only be partially filled with a rod of uniform cross-section.

Received July 26, 1957. The research reported in this document was supported jointly by the Army, Navy and Air Force under contract with Massachusetts Institute of Technology.

In this case, the first order variation of $(\gamma_0^2 - \gamma^2)$, the difference between the square of the propagation constant of the wave in the empty and loaded guide, is zero with respect to similar variations of the fields in the rod. Berk [3] has treated these problems obtaining variational principles which yield the differential equations satisfied by the electromagnetic field. His expressions, however, are not applicable to lossy media in general, and furthermore require a knowledge of the fields everywhere within the guide. Our method in addition provides us with a systematic way of improving the trial wave function, thus permitting us to improve upon the results obtained by perturbation formulae.

While the full use of the variational method consists of the substitution of trial functions with unknown variational parameters followed by the calculation of the stationary quantity, we may at times obtain good first order results by simply substituting completely determined trial functions. In such cases the variational expression reduces essentially to a perturbation formula [4]. It has the advantage, however, of furnishing the best amplitude of the trial field within the perturbing rod.

I. Integral equations for the electromagnetic field. The problem of a waveguide partially or completely filled by a rod of uniform cross-section permits separation with respect to the coordinate of the guide axis which we choose as the z -axis. Assuming an $\exp(j\omega t)$ time dependence, we can write the electric and magnetic fields as

$$\mathbf{E}(x, y) \exp[j(\omega t + \gamma z)]$$

and

$$\mathbf{H}(x, y) \exp[j(\omega t + \gamma z)]$$

respectively, where γ , the propagation constant, is in general a complex quantity.

In terms of $\mathbf{E}(x, y)$ and $\mathbf{H}(x, y)$, the z independent parts of the electromagnetic field, Maxwell's equations in a region free of charges and currents reduce to

$$\begin{aligned} (\nabla_i + j\gamma\mathbf{k}) \times \mathbf{E} &= -j\omega\mathbf{u} \cdot \mathbf{H}, \\ (\nabla_i + j\gamma\mathbf{k}) \times \mathbf{H} &= j\omega\mathbf{e} \cdot \mathbf{E}. \end{aligned} \quad (1)$$

\mathbf{e} is the tensor electric permittivity, and \mathbf{u} is the tensor magnetic permeability of the medium filling the guide. In the usual manner, by taking the curl of Maxwell's equations one obtains the wave equations satisfied by \mathbf{E} and \mathbf{H} . We find

$$\mathbf{d} \times \mathbf{d} \times \mathbf{E} - \omega^2 \epsilon_0 \mu_0 \mathbf{E} = \omega^2 \mu_0 (\mathbf{e}' \cdot \mathbf{E}) - j\omega \mathbf{d} \times (\mathbf{u}' \cdot \mathbf{H}), \quad (2a)$$

and

$$\mathbf{d} \times \mathbf{d} \times \mathbf{H} - \omega^2 \epsilon_0 \mu_0 \mathbf{H} = \omega^2 \epsilon_0 (\mathbf{u}' \cdot \mathbf{H}) + j\omega \mathbf{d} \times (\mathbf{e}' \cdot \mathbf{E}). \quad (2b)$$

We have set

$$\mathbf{d} = \nabla_i + j\gamma\mathbf{k},$$

$$\mathbf{e} = \epsilon_0 \mathbf{I} + \mathbf{e}',$$

$$\mathbf{u} = \mu_0 \mathbf{I} + \mathbf{u}',$$

where \mathbf{I} is the unit dyadic, and ϵ_0 and μ_0 are scalars whose choice depends on the problem at hand.

From the form of Maxwell's equations, we realize that the boundary conditions on

\mathbf{E} , \mathbf{H} , \mathbf{B} , and \mathbf{D} are not affected by the introduction of tensor ϵ and μ . Across a surface where ϵ or μ or both change discontinuously, the tangential components of \mathbf{E} and \mathbf{H} , and the normal components of \mathbf{B} and \mathbf{D} are continuous. At a conducting surface we shall require the tangential component of \mathbf{E} and the normal component of \mathbf{B} to be zero.

At this point we should also like to introduce the adjoint fields, \mathbf{E}^a and \mathbf{H}^a , which we take as the solutions of the equations

$$\begin{aligned}\mathbf{d}^a \times \mathbf{E}^a &= j\omega \mathbf{H}^a \cdot \mathbf{y} = j\omega \mathbf{B}^a, \\ \mathbf{d}^a \times \mathbf{H}^a &= -j\omega \mathbf{E}^a \cdot \epsilon = -j\omega \mathbf{D}^a,\end{aligned}$$

where $\mathbf{d}^a = \nabla_i - j\gamma \mathbf{k}$.

In the case of lossless media where ϵ and μ are hermitian [5], $\mathbf{E}^a = \mathbf{E}^*$ and $\mathbf{H}^a = \mathbf{H}^*$. In the case of lossy media, the adjoint fields may also be simply related to the fields \mathbf{E} and \mathbf{H} . For example in the case of lossy ferrites, taking Lax's form of the magnetic permeability tensor [4], having its diagonal elements even in ω and its off diagonal elements odd in ω ,

$$\mathbf{H}^a(x, y, \gamma, \omega) = \mathbf{H}(x, y, -\gamma, -\omega),$$

and

$$\mathbf{E}^a(x, y, \gamma, \omega) = \mathbf{E}(x, y, -\gamma, -\omega).$$

Consider now the problem of a waveguide of cross-section ($S_1 + S_2$) with region S_1 occupied by a material having tensor electromagnetic properties. We are confronted with the problem of solving Eqs. (1) within the two regions subject to the boundary conditions stated above. To do so, we* employ the help of the dyadic Green's function which satisfies the differential equation

$$\mathbf{d}^a \times \mathbf{d}^a \times \mathbf{Z}^a(\mathbf{x}|\mathbf{x}') - \omega^2 \epsilon_0 \mu_0 \mathbf{Z}^a(\mathbf{x}|\mathbf{x}') = \mathbf{I} \delta(\mathbf{x} - \mathbf{x}')$$
 (3)

everywhere in the guide, where $\mathbf{x} = xi + yj$.

The Dirac delta function $\delta(\mathbf{x} - \mathbf{x}') = \delta(x - x')\delta(y - y')$. We further require \mathbf{Z}^a to satisfy certain boundary conditions at the boundary of the guide. The Green's function $\mathbf{N}^a(\mathbf{x}|\mathbf{x}')$, which satisfies the condition

$$\mathbf{n} \times [\mathbf{N}^a(\mathbf{x}|\mathbf{x}')] = 0$$

when \mathbf{x} lies on the boundary of the guide, we refer to as the electric dyadic Green's function.

Analogously the dyadic Green's function, $\mathbf{M}^a(\mathbf{x}|\mathbf{x}')$, which satisfies the condition

$$\mathbf{n} \times [\mathbf{d}^a \times \mathbf{M}^a(\mathbf{x}|\mathbf{x}')] = 0,$$

when \mathbf{x} lies on the boundary of the guide, we call the magnetic dyadic Green's function.

For the solution of the fields in problems where either \mathbf{E} or \mathbf{H} is divergenceless, we further require the respective dyadic Green's function to satisfy the condition

$$\mathbf{d}^a \cdot \mathbf{Z}^a(\mathbf{x}|\mathbf{x}') = 0.$$

We now rewrite Eq. (3) in the form

$$\nabla_i \times (\mathbf{d}^a \times \mathbf{N}_a^a) - j\gamma \mathbf{k} \times (\mathbf{d}^a \times \mathbf{N}_a^a) - k_0^2 \mathbf{N}_a^a = \alpha \delta(\mathbf{x} - \mathbf{x}'),$$

where we have set $k_0^2 = \omega^2 \epsilon_0 \mu_0$ and $\mathbf{N}_a^a = \mathbf{N}^a \cdot \alpha$. α is an arbitrary vector introduced in

*We follow closely the work of J. Schwinger [2] on the problem of obstacles in cavities.

order to simplify the handling of the dyadic Green's function. Taking the dot product of Eq. (2a) with \mathbf{N}_α^a , the above equation with \mathbf{E} and subtracting, we find that

$$\begin{aligned}\mathbf{E} \cdot \alpha \delta(\mathbf{x} - \mathbf{x}') &= \nabla_{\mathbf{x}'} \cdot [\mathbf{N}_\alpha^a \times (\mathbf{d} \times \mathbf{E}) - \mathbf{E} \times (\mathbf{d}^a \times \mathbf{N}_\alpha^a) + j\omega \mathbf{N}_\alpha^a \times (\mathbf{u}' \cdot \mathbf{H})] \\ &\quad + \omega^2 \mu_0 (\epsilon' \cdot \mathbf{E}) \cdot \mathbf{N}_\alpha^a - j\omega (\mathbf{d}^a \times \mathbf{N}_\alpha^a) \cdot (\mathbf{u}' \cdot \mathbf{H}).\end{aligned}$$

Utilization of Maxwell's equations and the boundary conditions satisfied by the tangential components of \mathbf{E} and \mathbf{H} , leads to*

$$\begin{aligned}\mathbf{E}(\mathbf{x}') \cdot \alpha &= \omega^2 \mu_0 \int \mathbf{N}_\alpha^a(\mathbf{x}|\mathbf{x}') \cdot \epsilon' \cdot \mathbf{E}(\mathbf{x}) dS - j\omega \int [\mathbf{d}^a \times \mathbf{N}_\alpha^a(\mathbf{x}|\mathbf{x}')] \cdot [\mathbf{u}' \cdot \mathbf{H}(\mathbf{x})] dS. \quad (4) \\ \int \nabla_{\mathbf{x}'} \cdot \mathbf{P}(x, y) dx dy &= \oint \mathbf{n} \cdot \mathbf{P} ds.\end{aligned}$$

The magnetic field may be computed directly from the previous equation through the use of Eq. (1), or it may be obtained by utilizing Eq. (2b) and the magnetic dyadic Green's function. By repeating the steps which led to Eq. (4), we obtain

$$\mathbf{H}(\mathbf{x}') \cdot \alpha = \omega^2 \epsilon_0 \int \mathbf{M}_\alpha^a \cdot (\mathbf{u}' \cdot \mathbf{H}) dS + j\omega \int (\mathbf{d}^a \times \mathbf{M}_\alpha^a) \cdot (\epsilon' \cdot \mathbf{E}) dS. \quad (5)$$

With the help of the reciprocity relations satisfied by the Green's functions it is readily shown that the two methods for obtaining \mathbf{H} are equivalent [6].

Utilizing the two dimensional Green's second identity

$$\begin{aligned}\int [\mathbf{A}^a \cdot (\mathbf{d} \times \mathbf{d} \times \mathbf{B}) - \mathbf{B} \cdot (\mathbf{d}^a \times \mathbf{d}^a \times \mathbf{A}^a)] dS \\ = \oint \mathbf{n} \cdot [\mathbf{B} \times (\mathbf{d}^a \times \mathbf{A}^a) - \mathbf{A}^a \times (\mathbf{d} \times \mathbf{B})] ds, \quad (6)\end{aligned}$$

we find upon setting

$$\mathbf{A}^a = \mathbf{Z}_\alpha^a(\mathbf{x}|\mathbf{x}')$$

and

$$\mathbf{B} = \mathbf{Z}_\beta(\mathbf{x}|\mathbf{x}''),$$

that

$$\mathbf{Z}(\mathbf{x}''|\mathbf{x}') = [\mathbf{Z}^a(\mathbf{x}'|\mathbf{x}'')]^T,$$

where \mathbf{Z}^+ is the transpose of \mathbf{Z} . Similarly, setting

$$\mathbf{A}^a = \mathbf{d}^a \times \mathbf{M}_\alpha^a(\mathbf{x}|\mathbf{x}')$$

and

$$\mathbf{B} = \mathbf{N}_\beta(\mathbf{x}|\mathbf{x}''),$$

we obtain

$$\mathbf{d}^a \times \mathbf{M}^a(\mathbf{x}''|\mathbf{x}') = [\mathbf{d} \times \mathbf{N}(\mathbf{x}'|\mathbf{x}'')]^T.$$

II. The normal mode expansion of the dyadic Green's function.

Our next task is

*We have used the two dimensional Green's theorem

that of finding an explicit expression for the dyadic Green's functions. Such an expression can, of course, take the form of an expansion in terms of the complete set of vector eigenfunctions satisfying the homogeneous equation

$$\mathbf{d}^a \times \mathbf{d}^a \times \mathbf{A}_n^a - (\kappa_n^2 + \gamma^2) \mathbf{A}_n^a = 0, \quad (7)$$

the boundary condition

$$\mathbf{n} \times \mathbf{A}_n^a = 0$$

for the electric modes, and the boundary condition

$$\mathbf{n} \times (\mathbf{d}^a \times \mathbf{A}_n^a) = 0$$

for the magnetic modes.

We can readily find such a set of vector eigenfunctions in terms of the usual normal modes of the guide. If E_{zn} and H_{zn} are the z independent parts of the z components of the usual TM and TE modes of the guide, we find that the vectors

$$\mathbf{F}_n^a(x, y, \gamma, \gamma_n, \omega) = \begin{cases} \frac{j\omega\mu_0}{\gamma_n^2 - \omega^2\epsilon_0\mu_0} \mathbf{k} \times \nabla_t H_{zn}^a \\ \left[\frac{-j\gamma_n}{\gamma_n^2 - \omega^2\epsilon_0\mu_0} \nabla_t - \frac{\gamma_n}{\gamma} \mathbf{k} \right] E_{zn}^a, \end{cases} \quad (8)$$

where

$$E_{zn}^a(x, y) = E_{zn}^*(x, y), \quad H_{zn}^a(x, y) = H_{zn}^*(x, y),$$

and

$$\mathbf{F}_n^a(x, y, \gamma, \gamma_n, \omega) = \begin{cases} \frac{-j\omega\mu_0}{\gamma_n^2 - \omega^2\epsilon_0\mu_0} \mathbf{k} \times \nabla_t H_{zn}^a \\ \left[\frac{j\gamma_n}{\gamma_n^2 - \omega^2\epsilon_0\mu_0} \nabla_t - \frac{\gamma_n}{\gamma} \mathbf{k} \right] E_{zn}^a \end{cases}$$

satisfy Eq. (6) and the boundary conditions which have to be satisfied by the electric modes. In addition they are divergenceless,

$$\mathbf{d}^a \cdot \mathbf{F}_n^a = 0.$$

For completeness, we need another set of vector eigenfunctions whose curl is zero,

$$\mathbf{d}^a \times \mathbf{f}_n^a = 0.$$

Such a function, which is the gradient of a scalar,

$$\mathbf{f}_n^a = \mathbf{d}^a \phi_n^a,$$

will satisfy Eq. (6) for

$$\kappa_n^2 = -\gamma^2,$$

and the boundary condition $\mathbf{n} \times \mathbf{f}_n^a = 0$, if $\phi_n^a = 0$ on the boundary. In order however, for the functions \mathbf{f}_n^a to form an orthogonal set among themselves, (they are automatically orthogonal to the functions \mathbf{F}_n^a), we choose for the set of scalar functions ϕ_n^a , the set satisfying the equation

$$(\mathbf{d}^a \cdot \mathbf{d}^a + \omega_n^2) \phi_n^a = 0.$$

Analogously we obtain the magnetic vector eigenfunctions, the divergenceless vector eigenfunctions,

$$\mathbf{G}_n^a(x, y, z, \gamma, \gamma_n, \omega) = \begin{cases} \left[\frac{-j\gamma_n}{\gamma_n^2 - \omega^2 \epsilon_0 \mu_0} \nabla_t - \frac{\gamma_n}{\gamma} \mathbf{k} \right] H_{zn}^a, \\ \frac{-j\omega \epsilon_0}{\gamma_n^2 - \omega^2 \epsilon_0 \mu_0} \mathbf{k} \times \nabla_t E_{zn}^a, \end{cases} \quad (9)$$

and the curl-less eigenfunctions,

$$\mathbf{g}_n^a = \mathbf{d}^a \psi_n^a,$$

with $\mathbf{n} \cdot \mathbf{d}^2 \psi_n^a = 0$ on the boundary. The functions \mathbf{g}_n^a are orthogonal to each other and the functions \mathbf{G}_n if the scalar functions ψ_n^a satisfy the equation

$$(\mathbf{d}^a \cdot \mathbf{d}^a + \omega_n'^2) \psi_n^a = 0.$$

We now demonstrate the orthogonality of the vector eigenfunctions as expressed by

$$\int \mathbf{F}_n \cdot \mathbf{F}_m^a dS = \Lambda_n^2 \delta_{nm}, \quad (a)$$

$$\int \mathbf{f}_n \cdot \mathbf{f}_m^a dS = \Omega_n^2 \delta_{nm}, \quad (b)$$

$$\int \mathbf{f}_n \cdot \mathbf{F}_m^a dS = 0, \quad (c)$$

$$\int \mathbf{G}_n \cdot \mathbf{G}_m^a dS = \lambda_n^2 \delta_{nm}, \quad (d) \quad (10)$$

$$\int \mathbf{g}_n \cdot \mathbf{g}_m^a dS = \eta_n^2 \delta_{nm}, \quad (e)$$

$$\int \mathbf{G}_n \cdot \mathbf{g}_m^a dS = 0. \quad (f)$$

Equations (a), (c), (d) and (f) are readily obtainable through the use of Eq. (6), whose right hand side vanishes if \mathbf{A} and \mathbf{B} are both either electric or magnetic vector eigenfunctions. From Eq. (7) it then follows that

$$(\kappa_n^2 - \kappa_m^2) \int \mathbf{A}_n^a \cdot \mathbf{B}_m dS = 0.$$

Equations (b) and (e) can be shown to hold since for θ_n being any scalar function

$$\begin{aligned} \int (\mathbf{d} \theta_m) \cdot (\mathbf{d}^a \theta_n^a) dS &= \int \nabla_t \cdot (\theta_m \mathbf{d}^a \theta_n^a) dS - \int \theta_m (\mathbf{d}^a \cdot \mathbf{d}^a \theta_n^a) dS \\ &= \int \nabla_t \cdot [(\mathbf{d} \theta_m) \theta_n^a] dS - \int \theta_n^a \mathbf{d} \cdot \mathbf{d} \theta_m dS. \end{aligned}$$

If now θ_n satisfies the equation

$$\mathbf{d}^a \cdot \mathbf{d}^a \theta_n^a + \omega_n^2 \theta_n^a = 0$$

and the boundary condition satisfied either by ϕ_n or ψ_n , we have that

$$\begin{aligned} \int d\theta_m \cdot d^\circ \theta_n dS &= \omega_n^2 \int \theta_m \theta_n^\circ dS \\ &= \omega_m^2 \int \theta_m \theta_n^\circ dS. \end{aligned}$$

Thus if $\omega_n^2 \neq \omega_m^2$ the left hand side must be zero. We assume that degenerate eigenfunctions have been orthogonalized. It is readily shown that the divergenceless electric and magnetic vector eigenfunctions are related through the equations

$$\mathbf{d} \times \mathbf{F}_n = -j\omega\mu_0 \mathbf{G}_n \alpha_n \quad (11)$$

and

$$\mathbf{d} \times \mathbf{G}_n = j\omega\epsilon_0 \mathbf{F}_n \beta_n,$$

where

$$\alpha_n = \begin{cases} \frac{\gamma}{\gamma_n} & \text{for TE modes} \\ \frac{\gamma_n k_0^2 + \gamma^2 - \gamma_n^2}{\gamma k_0^2} & \text{for TM modes} \end{cases}$$

and

$$\beta_n = \begin{cases} \frac{\gamma_n k_0^2 + \gamma^2 - \gamma_n^2}{\gamma k_0^2} & \text{for TE modes} \\ \frac{\gamma}{\gamma_n} & \text{for TM modes.} \end{cases}$$

The functions \mathbf{F}_n and \mathbf{f}_n together form a complete set of orthogonal vector eigenfunctions which may be used to represent any vector field having zero tangential component on the boundary of the guide. Similarly the set of orthogonal vector eigenfunctions \mathbf{G}_n and \mathbf{g}_n may be used to represent any vector field satisfying the same boundary conditions as those imposed upon the magnetic vector eigenfunctions. We should thus be able to write

$$\mathbf{N}_a(\mathbf{x}|\mathbf{x}') = \sum_n a_n \mathbf{F}_n + \sum_n b_n \mathbf{f}_n, \quad (12)$$

and

$$\mathbf{M}_a(\mathbf{x}|\mathbf{x}') = \sum_n a'_n \mathbf{G}_n + \sum_n b'_n \mathbf{g}_n. \quad (13)$$

Realizing that the normal modes satisfy Eq. (2) with ϵ' and \mathbf{u}' equal to zero and $k_0^2 = \omega^2 \epsilon_0 \mu_0$ replaced by $(\kappa_n^2 + \gamma^2)$, we find

$$\mathbf{A}_n(\mathbf{x}') \cdot \boldsymbol{\alpha} = (\kappa_n^2 + \gamma^2 - k_0^2) \int \mathbf{A}_n(\mathbf{x}) \cdot \mathbf{Z}_a(\mathbf{x}|\mathbf{x}') dS$$

as a solution of Eq. (7). Inserting expansions (12) and (13) for the dyadic Green's functions into this integral equation, and utilizing the orthogonality relations (10), we find

that

$$\mathbf{a}_n = \frac{\mathbf{F}_n(\mathbf{x}') \cdot \boldsymbol{\alpha}}{(\gamma^2 - \gamma_n^2) \lambda_n^2},$$

$$\mathbf{b}_n = \frac{-\mathbf{f}_n(\mathbf{x}') \cdot \boldsymbol{\alpha}}{k_0^2 \Omega_n^2},$$

$$\mathbf{a}'_n = \frac{\mathbf{G}_n(\mathbf{x}') \cdot \boldsymbol{\alpha}}{(\gamma^2 - \gamma_n^2) \lambda_n^2},$$

and

$$\mathbf{b}'_n = -\frac{\mathbf{g}_n(\mathbf{x}') \cdot \boldsymbol{\alpha}}{k_0^2 \eta_n^2}.$$

We have thus found the expansion of the dyadic Green's function in terms of the normal modes,

$$\mathbf{N}''(\mathbf{x}|\mathbf{x}') = \sum_n \frac{\mathbf{F}_n''(\mathbf{x}) \mathbf{F}_n(\mathbf{x}')}{(\gamma^2 - \gamma_n^2) \lambda_n^2} - \frac{1}{k_0^2} \sum_n \frac{\mathbf{f}_n''(\mathbf{x}) \mathbf{f}_n(\mathbf{x}')}{\Omega_n^2}, \quad (14)$$

and

$$\mathbf{M}''(\mathbf{x}|\mathbf{x}') = \sum_n \frac{\mathbf{G}_n''(\mathbf{x}) \mathbf{G}_n(\mathbf{x}')}{(\gamma^2 - \gamma_n^2) \lambda_n^2} - \frac{1}{k_0^2} \sum_n \frac{\mathbf{g}_n''(\mathbf{x}) \mathbf{g}_n(\mathbf{x}')}{\eta_n^2}. \quad (15)$$

For those problems where \mathbf{N} or \mathbf{M} is taken to be divergenceless, the sum over the \mathbf{f}_n 's or \mathbf{g}_n 's respectively, except for the terms which are also divergenceless, are not to be included.

III. Variational principles. In terms of the normal mode expansion of the dyadic Green's function

$$\mathbf{E}(\mathbf{x}') = \sum_{n=0}^{\infty} \frac{J_n'' \mathbf{E}_n(\mathbf{x}')}{(\gamma^2 - \gamma_n^2)} - \frac{\omega^2 \mu_0}{k_0^2} \sum_{n=0}^{\infty} \frac{\left[\int \mathbf{f}_n'' \cdot \boldsymbol{\epsilon}' \cdot \mathbf{E} dS \right] \mathbf{f}_n(\mathbf{x}')}{\Omega_n^2}, \quad (16)$$

where having set

$$\mathbf{E}_n = (\beta_n)^{1/2} \mathbf{F}_n,$$

and

$$\begin{aligned} \mathbf{H}_n &= (\alpha_n)^{1/2} \mathbf{G}_n, \\ J_n'' &= \frac{\omega^2 \mu_0}{\beta_n \lambda_n^2} \left[\int \mathbf{E}_n'' \cdot \boldsymbol{\epsilon}' \cdot \mathbf{E} dS + (\alpha_n \beta_n)^{1/2} \int \mathbf{H}_n'' \cdot \mathbf{u}' \cdot \mathbf{H} dS \right]. \end{aligned} \quad (17)$$

Taking the curl of Eq. (16) yields

$$\frac{\mathbf{u}' \cdot \mathbf{H}(\mathbf{x}')}{\mu_0} = \sum_{n=0}^{\infty} \frac{(\alpha_n \beta_n)^{1/2} J_n''}{(\gamma^2 - \gamma_n^2)} \mathbf{H}_n(\mathbf{x}') - \mathbf{H}(\mathbf{x}'). \quad (18)$$

We note again, that in those problems where \mathbf{E} is divergenceless we shall omit the last sum in Eq. (16), except for the \mathbf{f}_0 term which is also divergenceless.

Analogous to the foregoing we find

$$\mathbf{E}^a(\mathbf{x}) = \sum_{n=0}^{\infty} \frac{(J_n^e)^a \mathbf{E}_n^a(\mathbf{x})}{(\gamma^2 - \gamma_n^2)} - \frac{\omega^2 \mu_0}{k_0^2} \sum_{n=0}^{\infty} \frac{\int \mathbf{E}^a \cdot \mathbf{e}' \cdot \mathbf{f}_n dS}{\Omega_n^2} \mathbf{f}_n^a(\mathbf{x}) \quad (19)$$

and

$$\frac{\mathbf{H}^a(\mathbf{x}) \cdot \mathbf{u}'}{\mu_0} = \sum_{n=0}^{\infty} \frac{(\alpha_n \beta_n)^{1/2} (J_n^e)^a \mathbf{H}_n^a(\mathbf{x})}{(\gamma^2 - \gamma_n^2)} - \mathbf{H}^a(\mathbf{x}). \quad (20)$$

From the wave equations satisfied by the electromagnetic field and the normal modes, we obtain

$$\begin{aligned} \nabla_t \cdot [\mathbf{E} \times \mathbf{d}^a \times \mathbf{E}_n^a + (\mathbf{d} \times \mathbf{E}) \times \mathbf{E}_n^a - j\omega \mathbf{E}_n^a \times (\mathbf{u}' \cdot \mathbf{H})] + (\gamma^2 - \gamma_n^2) \mathbf{E}_n^a \cdot \mathbf{E} \\ = \omega^2 \mu_0 \mathbf{E}_n^a \cdot \mathbf{e}' \cdot \mathbf{E} - j\omega (\mathbf{d}^a \times \mathbf{E}_n^a) \cdot (\mathbf{u}' \cdot \mathbf{H}). \end{aligned}$$

It follows therefore that

$$J_n^e = (\gamma^2 - \gamma_n^2) \int \mathbf{E}_n^a \cdot \mathbf{E} dS / \beta_n \Lambda_n^2. \quad (21)$$

We can fix the amplitude of the field within the rod by restricting ourselves to fields satisfying the condition $\int \mathbf{E}_0^a \cdot \mathbf{E} dS = (\beta_0)^{1/2} \Lambda_0^2$, where E_0 is the unperturbed guide solution whose perturbation we seek. Thus,

$$J_0^e = (\gamma^2 - \gamma_0^2) (\beta_0)^{-1/2}, \quad (22)$$

and utilizing Eqs. (16) and (18), it follows that

$$(J_0^e)^a (\beta_0)^{-1/2} \Lambda_0^2 = \omega^2 \mu_0 D_0^e, \quad (23)$$

where

$$\begin{aligned} D_0^e = \int \mathbf{E}^a \cdot \mathbf{e}' \cdot \mathbf{E} dS + \frac{1}{\mu_0} \int \mathbf{H}^a \cdot \mathbf{u}' \cdot \mathbf{u}' \cdot \mathbf{H} dS + \int \mathbf{H}^a \cdot \mathbf{u}' \cdot \mathbf{H} dS \\ - \frac{1}{\omega^2 \mu_0} \sum_{n=1}^{\infty} \frac{\beta_n \Lambda_n^2 J_n^e (J_n^e)^a}{(\gamma^2 - \gamma_n^2)} \\ + \frac{\omega^2 \mu_0}{k_0^2} \sum_{n=0}^{\infty} \left[\int \mathbf{f}_n^a \cdot \mathbf{e}' \cdot \mathbf{E} dS \right] \left[\int \mathbf{E}^a \cdot \mathbf{e}' \cdot \mathbf{f}_n dS \right]. \end{aligned} \quad (24)$$

We can now set up a variational principle from which integral equations (16), (18), (19), and (20) are derivable. Combining Eqs. (22) and (23) we obtain the expression

$$\frac{(\gamma^2 - \gamma_0^2) \omega^2 \mu_0}{\Lambda_0^2} = \frac{J_0^e (J_0^e)^a}{D_0^e}. \quad (25)$$

It is readily shown that the first order variation of the right hand side of Eq. (25) with respect to the fields, subject to the condition

$$J_0^e = (J_0^e)^a = D_0^e \omega^2 \mu_0 (\beta_0)^{1/2} \Lambda_0^{-2} \quad (26)$$

yields the integral equations for the electromagnetic fields within the rod. Since expression (25) is amplitude independent, this condition will be automatically satisfied.

Quite analogous to the foregoing we can obtain another variational principle yielding the integral equations one would obtain starting with Eq. (5). These integral equations

are

$$\mathbf{H}(\mathbf{x}') = \sum_n \frac{J_n^m \mathbf{H}_n(\mathbf{x}')}{(\gamma^2 - \gamma_n^2)} - \frac{\omega^2 \epsilon_0}{k_0^2} \sum_n \left[\int \mathbf{g}_n^a \cdot \mathbf{v}' \cdot \mathbf{H} dS \right] \frac{1}{\eta_n^2} \mathbf{g}_n(\mathbf{x}'), \quad (27)$$

$$\frac{\mathbf{e}' \cdot \mathbf{E}(\mathbf{x}')}{\epsilon_0} = \sum_n \frac{(\alpha_n \beta_n)^{1/2} J_n^m \mathbf{E}_n(\mathbf{x}')}{(\gamma^2 - \gamma_n^2)} - \mathbf{E}(\mathbf{x}'), \quad (28)$$

and their adjoints.

$$\begin{aligned} J_n^m &= \frac{\omega^2 \epsilon_0}{\alpha_n \lambda_n^2} \left[\int \mathbf{H}_n^a \cdot \mathbf{v}' \cdot \mathbf{H} dS + (\alpha_n \beta_n)^{1/2} \int \mathbf{E}_n^a \cdot \mathbf{e}' \cdot \mathbf{E} dS \right] \\ &= (\gamma^2 - \gamma_n^2) \int \mathbf{H}_n^a \cdot \mathbf{H} dS / \alpha_n \lambda_n^2. \end{aligned} \quad (29)$$

Restricting ourselves to solutions which satisfy the condition $\int \mathbf{H}_0^a \cdot \mathbf{H} dS = (\alpha_0)^{1/2} \lambda_0^2$, we find

$$J_0^m = (\gamma^2 - \gamma_0^2) (\alpha_0)^{-1/2} \quad (30)$$

and

$$(J_0^m)^a (\alpha_0)^{-1/2} \lambda_0^2 = \omega^2 \epsilon_0 D_0^m, \quad (31)$$

where

$$\begin{aligned} D_0^m &= \frac{1}{\epsilon_0} \int \mathbf{E}^a \cdot \mathbf{e}' \cdot \mathbf{e}' \cdot \mathbf{E} dS + \int \mathbf{E}^a \cdot \mathbf{e}' \cdot \mathbf{E} dS + \int \mathbf{H}^a \cdot \mathbf{v}' \cdot \mathbf{H} dS \\ &\quad - \frac{1}{\omega^2 \epsilon_0} \sum_{n=1}^{\infty} \frac{\alpha_n \lambda_n^2 J_n^m (J_n^m)^a}{(\gamma^2 - \gamma_n^2)} \\ &\quad + \frac{\omega^2 \epsilon_0}{k_0^2} \sum_{n=0}^{\infty} \left[\int \mathbf{g}_n^a \cdot \mathbf{v}' \cdot \mathbf{H} dS \right] \left[\int \mathbf{H}^a \cdot \mathbf{v}' \cdot \mathbf{g}_n dS \right]. \end{aligned}$$

The variational expression from which these integral equations for the fields within the rod are derivable is

$$\frac{(\gamma^2 - \gamma_0^2) \omega^2 \epsilon_0}{\lambda_0^2} = \frac{J_0^m (J_0^m)^a}{D_0^m}. \quad (32)$$

IV. Application of the variational principle. In the present application of the variational principle, we are interested in demonstrating its usefulness in improving the results obtained by perturbation methods. Consider the problem of a rectangular wave-

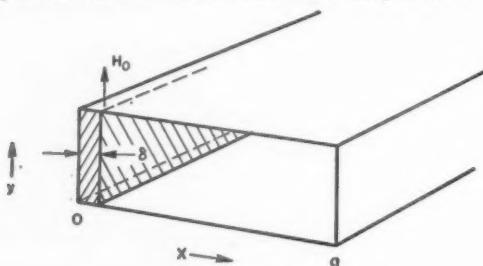


FIG. 1. A single ferrite slab in a rectangular waveguide.

guide containing a thin ferrite slab, transversely magnetized and placed against one of the sides of the guide (Fig. 1). The exact solution of the problem [7] leads to a complicated transcendental equation for the evaluation of the propagation constant. We shall obtain the first order perturbation of the propagation constant of the lowest propagating mode of the guide, the TE_{10} mode, for which

$$(\mathbf{F}_0) = \begin{bmatrix} 0 \\ -\frac{j\omega\mu_0}{\pi/L} \sin \frac{\pi x}{L} \\ 0 \end{bmatrix} \quad \text{and} \quad (\alpha_0 \mathbf{G}_0) = \begin{bmatrix} -\frac{j\gamma}{\pi/L} \sin \frac{\pi x}{L} \\ 0 \\ \cos \frac{\pi x}{L} \end{bmatrix}.$$

The transversely magnetized ferrite slab has a tensor magnetic permeability

$$(\boldsymbol{\mu}) = \mu_0 \begin{bmatrix} \mu_1 & 0 & j\kappa \\ 0 & 1 & 0 \\ -j\kappa & 0 & \mu_1 \end{bmatrix},$$

where μ_1 and κ are functions of the frequency and the external DC magnetic field [8]. The electric permittivity, ϵ , is a scalar, and \mathbf{E} therefore divergenceless.

For first order results, the choice

$$\mathbf{E}_{\text{trial}} = A \mathbf{F}_0,$$

where A is a constant to be determined, is as good as any other. Its use in the perturbation formula

$$(\gamma^2 - \gamma_0^2) \frac{\Lambda_0^2}{\omega^2 \mu_0} = \int \mathbf{F}_0 \cdot \mathbf{e}' \cdot \mathbf{E} dx + \int \alpha_0 \mathbf{G}_0 \cdot \mathbf{y}' \cdot \mathbf{H} dx = j_0 \quad (33)$$

which can be obtained from Eqs. (17) and (22), and is equivalent to perturbation formula (7) of [4], leads to the result

$$\frac{(\gamma^2 - \gamma_0^2) \Lambda_0^2}{\omega^2 \mu_0} = A \left\{ \frac{\mu_0^2 \epsilon'}{(\pi/L)^2} d_{11} + \frac{\mu_0}{\omega^2} \left[\left(\frac{\mu_1^2 - \mu_1 - \kappa^2}{\mu_1^2 - \kappa^2} \right) d_{22} \right] \right. \\ \left. - \frac{\mu_0}{\omega^2} \left[\frac{2\kappa}{(\mu_1^2 - \kappa^2)} \frac{\gamma}{\pi/L} d_{12} \right] \right\}, \quad (34)$$

where

$$d_{11} = \int_0^L \sin^2 \frac{\pi x}{L} dx, \quad d_{12} = \int_0^L \sin \frac{\pi x}{L} \cos \frac{\pi x}{L} dx, \\ d_{22} = \int_0^L \cos^2 \frac{\pi x}{L} dx, \quad \text{and} \quad \Lambda_0^2 = \frac{\omega^2 \mu_0}{(\pi/L)^2} \frac{L}{2}.$$

We have set

$$\mathbf{H}_{\text{trial}} = \frac{\mu_1^2 - \mu_1 - \kappa^2}{\mu_1^2 - \kappa^2} \mu_0 \alpha_0 \mathbf{G}_0 - \frac{j\mu_0}{\mu_1^2 - \kappa^2} (\mathbf{j} \times \mathbf{G}_0) \alpha_0$$

and have kept only those terms which yield terms of the second order in δ or less, except

for large values of ϵ , where in the expressions containing ϵ we have kept terms yielding terms of the third order in δ or less. We have also neglected the integrations with respect to the y coordinate as they all yield the same value or zero.

We find the perturbation result to be dependent on the amplitude of the trial field. In this problem for which an exact solution exists, one is able to obtain an estimate of the amplitude. In general, however, we are not as fortunate, and in such cases Eq. (23) may be used to obtain a first order approximation of the amplitude. Setting $\mathbf{E} = A\mathbf{e}$ and $\mathbf{H} = Ah$ we find from Eq. (23) that

$$A = \frac{(j_0^e)^a \Lambda_0^2}{\omega^2 \mu_0 \mathfrak{D}_0^e}, \quad (35)$$

where

$$(j_0^e)^a = \int \mathbf{e}^a \cdot \mathbf{e}' \cdot \mathbf{F}_0 dS + \int \mathbf{h}^a \cdot \mathbf{h}' \cdot (\alpha_0 \mathbf{G}_0) dS$$

and

$$A^2 \mathfrak{D}_0^e = D_0^e.$$

Combination of Eqs. (33) and (35) yields the variational expression

$$\frac{\gamma^2 - \gamma_0^2}{\omega^2 \mu_0} \Lambda_0^2 = \frac{j_0^e (j_0^e)^a}{\mathfrak{D}_0^e}. \quad (36)$$

Before evaluating \mathfrak{D}_0^e let us note that since the fields are independent of the y coordinate, we can find a closed dyadic Green's function satisfying the differential equation

$$\mathbf{d}^a \times \mathbf{d}^a \times \mathbf{N}(x|x') - k_0^2 \mathbf{N}(x|x') = \mathbf{I}\delta(x - x'),$$

where

$$\mathbf{d}^a = \frac{\partial}{\partial x} \mathbf{i} - j\gamma \mathbf{k}.$$

Such a dyadic Green's function is given by

$$\mathbf{N}(x|x') = \mathbf{N}_1(x|x') + \mathbf{N}_2(x|x') - \frac{1}{k_0^2} (\mathbf{d}^a \mathbf{d}' \cdot \mathbf{N}_1 - \mathbf{d}^a \mathbf{d}' \cdot \mathbf{N}_2),$$

$$\mathbf{N}_1(x|x') = \mathbf{I}\mathfrak{N}_1(x|x')$$

where the scalar Green's function

$$\mathfrak{N}_1(x|x') = -\frac{1}{2k \sin kL} \begin{cases} \cos k(x - x' - L) & x > x' \\ \cos k(x - x' + L) & x < x' \end{cases}$$

and satisfies the differential equation

$$\frac{d^2}{dx^2} \mathfrak{N}_1 + k^2 \mathfrak{N}_1 = -\delta(x - x').$$

$$\mathbf{N}_2(x|x') = -ii\mathfrak{N}_2 + jj\mathfrak{N}_2 + kk\mathfrak{N}_2$$

where

$$\mathfrak{N}_2(x|x') = \frac{1}{2k \sin kL} \cos k(x + x' - L)$$

and

$$k^2 = k_0^2 - \gamma^2.$$

In terms of this dyadic Green's function

$$\begin{aligned} \mathfrak{D}_0^* &= \int \mathbf{e}^a \cdot \mathbf{e}' \cdot \mathbf{e} dx + \int \mathbf{h}^a \cdot \mathbf{h}' \cdot \mathbf{h} dx + \frac{1}{\mu_0} \int \mathbf{h}^a \cdot \mathbf{y}' \cdot \mathbf{y}' \cdot \mathbf{h} dx \\ &\quad - \iint \mathbf{e}^a \cdot \mathbf{e}' \cdot \mathbf{N}^T(x|x') \cdot \mathbf{e}' \cdot \mathbf{e} dx dx' \\ &\quad - \frac{1}{\mu_0} \iint (\mathbf{h}^a \cdot \mathbf{y}') \cdot \mathbf{d}' \times [\mathbf{d}^a \times \mathbf{N}(x|x')]^T \cdot (\mathbf{y}' \cdot \mathbf{h}) dx dx' \\ &\quad + j\omega \iint (\mathbf{e}^a \cdot \mathbf{e}') \cdot [\mathbf{d}^a \times \mathbf{N}(x|x')]^T \cdot (\mathbf{y}' \cdot \mathbf{h}) dx dx' \\ &\quad - j\omega \iint (\mathbf{h}^a \cdot \mathbf{y}') \cdot [\mathbf{d} \times \mathbf{N}^T(x|x')] \cdot (\mathbf{e}' \cdot \mathbf{e}) dx dx' \\ &\quad + \frac{\omega^2 \mu_0}{(\gamma^2 - \gamma_0^2) \Lambda_0^2} [j_0^*(j_0^*)^a]. \end{aligned}$$

Inserting the trial fields, we find that to first order

$$\mathfrak{D}_0^* = \frac{\mu_1}{\mu_1^2 - \kappa^2} (j_0^*)^a$$

or

$$A = \frac{\mu_1^2 - \kappa^2}{\mu_1}.$$

Thus

$$(\gamma^2 - \gamma_0^2) \frac{L}{2} = \omega^2 \mu_0 \epsilon' \left(\frac{\mu_1 - \kappa^2}{\mu_1} \right) d_{11} + \frac{(\mu_1^2 - \mu_1 - \kappa^2)}{\mu_1} \left(\frac{\pi}{L} \right)^2 d_{22} - \frac{2\kappa}{\mu_1} \left(\frac{\pi}{L} \right) \gamma d_{12}. \quad (37)$$

It might be of interest to note that the amplitude one finds utilizing Eq. (35) is the limit of the exact amplitude as δ becomes very small.

In Fig. 2 we plot γ for a wave at 9000 mcps as a function of δ , as computed from the above equation, for a ferrite slab having a magnetization of 3000 gauss placed in an external DC magnetic field yielding an internal value of 1000 oersteds, and compare the approximate results with the exact calculation of Lax and Button [7].

While the full use of the variational method consists of the substitution of trial functions with unknown variational parameters and finding the values of these parameters which make the variational expression an extremal, we see that we may at times obtain good first order results by simply substituting completely determined trial functions. In such cases the variational expression will yield better results than the perturbation formula with the same trial field.

Acknowledgment. The author would like to express his appreciation to his colleagues for their stimulation and helpful discussions. Specifically he is indebted to Drs. L. Gold, G. S. Heller, B. Lax, H. J. Zeiger and Mr. K. J. Button. The numerical work was done by Miss Clare M. Glennon.

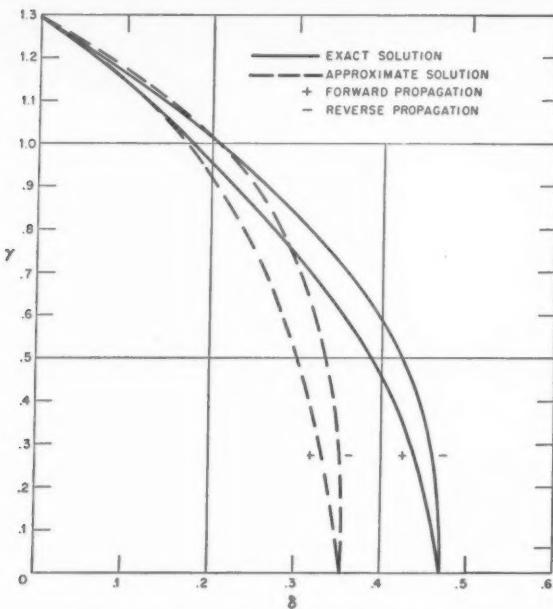


FIG. 2. Propagation constant for the TE_{10} guide mode perturbed by a single ferrite slab at 9000 mcps as a function of slab thickness.

REFERENCES

1. B. Lax and G. S. Heller, private communication. Suhl and Walker, *Topics in guided wave propagation through gyro-magnetic media Part III*, Bell System Tech. J. **XXXIII**, II33 (1954)
2. J. Schwinger, *The theory of obstacles in resonant cavities and waveguides*, M.I.T., Rad. Lab. Rept. 43-34 (May 1943)
3. A. D. Berk, *Variational principles for electromagnetic resonators and waveguides*, I.R.E., AP **4**, 104 (1956)
4. B. Lax, *Frequency and loss characteristics of microwave ferrite devices*, Proc. I. R. E. **44**, 1368 (1956)
5. B. D. H. Tellegen, *The synthesis of passive resistanceless four-poles that may violate the reciprocity relation*, Phillips Research Repts. **3**, 321 (1948)
6. W. Hauser, *Variational principles for guided electromagnetic waves in anisotropic materials*, M.I.T., Lincoln Lab. Rept. M35-61 (August 1956); not generally available
7. B. Lax and K. J. Button, *Theory of new ferrite modes in rectangular waveguide*, J. Appl. Phys. **26**, 1184 (1955)
8. D. Polder, Phil. Mag. **40**, 99 (1949)

ON TRANSFER FUNCTIONS AND TRANSIENTS*

BY

ARMEN H. ZEMANIAN
New York University

Abstract. In the first part of this paper the concept of the positive real function is generalized so that it is applicable to transfer functions and the functions, satisfying this generalized concept, are arranged into classes. Some tests are then developed which may be used to determine whether a transfer function belongs to a particular class. It is also shown that if transfer functions have certain general forms then they will automatically be members of one of the classes. Finally, several properties of the phase functions for such system functions are developed.

The second part of the paper considers the impulse and step responses corresponding to these transfer functions. It is found that these transient responses are bounded and moreover the rise time and settling time of the step response are found to be greater than lower bounds which depend on the amount of the maximum overshoot (or undershoot). These results are also generalizations of the restrictions developed on the transient responses of positive real system functions and are considerably stronger. As the difference in degree between the numerator and denominator of the transfer function increases, the magnitudes of these lower bounds also increase.

Introduction. The question of what restrictions exist on the transient responses of various classes of networks has been treated in a series of papers [1-3]. In particular, bounds on the impulse and step responses have been given when the corresponding frequency responses are restricted to being of constant sign or monotonic in a semi-infinite interval. An example of this is that, if a rational system function, $Z(s)$, is positive real and has one more pole than zero, then the absolute value of the impulse response is never greater than $1/C$ which is the constant multiplier of the system function. These results lead in turn to lower bounds on the rise time or settling time of the step response of such networks when the overshoot and undershoot is specified.

Recently, one of these results has been improved by Ovseyevich [4]. The improvement is on the bounds developed on the impulse response of the positive real system functions in the interval $(0, T)$ when the bounds on that response in the interval (T, ∞) are known. A related problem has been considered by Cuteridge [5] who determines the optimum rise time of passive two-terminal networks under various restrictions. For instance, considering only the step responses which are monotonic or have only a small overshoot, the optimum rise time is determined when the system function has only two zeros and three poles.

This paper deals with a generalization of these results which is applicable to transfer functions. In particular, certain classes of functions are defined from the system functions having any number of poles in excess of zeros such that the corresponding transient responses are restricted in a manner similar to the restrictions existing on the transient responses of positive real functions. The strength of the restriction on the transient responses increases as the excess of poles over zeros increases.

In the first part of this paper, these classes of system functions are defined. The

*Received September 3, 1957.

definition may be considered a generalization of the concept of positive real system functions which is suitable for application to transfer functions. Various tests are then developed which determine whether a system function is a member of any class. One immediate result is that, if the system function has poles only in the left hand complex frequency plane and no zeros anywhere except at infinity (that is, if the system function is the reciprocal of a Hurwitz polynomial), then it will be a member of one of the classes. Furthermore, certain properties of the phase functions for such system functions are also developed.

In the second section the restrictions on the transient responses for the defined classes of functions are derived. These constitute bounds on the impulse responses and the step responses. From these, lower bounds are developed on the rise time and settling time of the step response when the overshoot and undershoot are given and, conversely, the specification of the rise time or settling time fixes the lower bounds on the maximum overshoot or undershoot.

Part I. A generalization of the concept of positive real functions. The systems considered in this paper are lumped, linear, fixed, finite and stable systems so that the system functions¹ $Z(s)$ have the following rational form where s is the complex frequency variable $\sigma + j\omega$, the coefficients and the constant multiplier K are real numbers, and n and m are positive integers ($m > n$).

$$Z(s) = K \frac{s^n + a_{n-1}s^{n-1} + \cdots + a_0}{s^m + b_{m-1}s^{m-1} + \cdots + b_0} = K \frac{N(s)}{D(s)}. \quad (1)$$

The term "stable system" is taken to mean a system whose response will eventually become arbitrarily small once the input is removed. Thus the polynomial $D(s)$ is a Hurwitz polynomial all of whose roots have a negative (non-zero) real part.

The special classes of functions that are of interest here will be defined after a few preliminary remarks. Let

$$Z_q(s) = (-j)^q \int_{-\infty}^s ds_{q-1} \int_{-\infty}^{s_{q-1}} ds_{q-2} \cdots \int_{-\infty}^{s_1} Z(s_0) ds_0, \quad (2)$$

where the real parts of the complex variables $s_0, s_1, \dots, s_{q-1}, s$ are all non-negative and $q \leq m - n - 1$. Letting $s_q = \sigma_q + j\omega_q$, the real and imaginary parts, $R_q(\omega)$ and $I_q(\omega)$, of $Z_q(j\omega)$ may be obtained from the real and imaginary parts, $R(\omega)$ and $I(\omega)$, of $Z(j\omega)$ by (3) and (4).

$$Z_q(j\omega) = R_q(\omega) + jI_q(\omega)$$

$$R_q(\omega) = \int_{-\infty}^{\omega} d\omega_{q-1} \int_{-\infty}^{\omega_{q-1}} d\omega_{q-2} \cdots \int_{-\infty}^{\omega_1} R(\omega_0) d\omega_0 \quad (3)$$

$$I_q(\omega) = \int_{-\infty}^{\omega} d\omega_{q-1} \int_{-\infty}^{\omega_{q-1}} d\omega_{q-2} \cdots \int_{-\infty}^{\omega_1} I(\omega_0) d\omega_0. \quad (4)$$

It is evident from the following argument that these successive integrations may be performed $m - n - 1$ times. Since $Z(s)$ is analytic for $\sigma \geq 0$, the integral of $Z(s)$ between two given points will yield the same result for all paths between these points which do not enter the left half s plane. Furthermore, the inverse power series expansion of $Z(s)$,

¹These system functions may be impedances, admittances or transfer functions that are ratios of currents or voltages.

which holds for $|s| \geq M$, is

$$Z(s) = \sum_{\mu=m-n}^{\infty} \frac{K_{\mu}}{s^{\mu}}, \quad (5)$$

where $K_{m-n} = K$ and M is a positive number greater than the distance from the origin to the pole of $Z(s)$ farthest away from the origin. Since this series converges uniformly for $|s| \geq M$, it may be integrated term by term yielding a series which also converges uniformly in the same region. This process may be continued q times where $q \leq m - n - 1$. Thus the successive integrations of $Z(s)$ given by (2) yield unique functions $Z_q(s)$ which are all analytic for $\sigma_q \geq 0$.

Now the aforementioned classes of functions $Z(s)$ given by (1) are defined as follows.

Definition. $Z(s)$ will be called a class k function, where $k = m - n$, if one of the following inequalities holds for $-\infty < \omega < +\infty$. For $k = 2\nu + 1$ ($\nu = 0, 1, 2, \dots$),

$$(-1)^{\nu} R_{k-1}(\omega) \geq 0 \quad (6)$$

and, for $k = 2\nu$ ($\nu = 1, 2, 3, \dots$),

$$(-1)^{\nu+1} I_{k-1}(\omega) \geq 0. \quad (7)$$

The class k functions have the interesting property that they are related to the positive real functions in the region where they are defined (that is, in the right half s plane and on the imaginary axis).

Theorem 1. If the system function $Z(s)$ is a class k function, then $(-j)^{m-n-1} Z_{m-n-1}(s)$ is a positive real function for $\sigma \geq 0$ and the constant multiplier K is positive.

Proof. The inequalities (6) and (7) state that the real part of $(-j)^{m-n-1} Z_{m-n-1}(j\omega)$ is non-negative for all ω . Moreover $Z_{m-n-1}(s)$ is analytic for $\sigma \geq 0$ since $Z(s)$ is analytic in this region. Thus by the minimax theorem, the real part of $(-j)^{m-n-1} Z_{m-n-1}(s)$ is non-negative for $\sigma \geq 0$. Therefore to prove the first part of the theorem it need only be shown that $(-j)^{m-n-1} Z_{m-n-1}(s)$ is real for $s = \sigma \geq 0$. But

$$(-j)^{m-n-1} Z_{m-n-1}(\sigma) = (-1)^{m-n-1} \int_{\infty}^{\sigma} d\sigma_{m-n-2} \int_{\infty}^{\sigma_{m-n-2}} d\sigma_{m-n-3} \cdots \int_{\infty}^{\sigma_1} Z(\sigma_0) d\sigma_0$$

and the right hand side of this expression is a real function of a real variable.

Finally the series expansion of $Z(s)$ given by (5) may be integrated term by term q times according to (2) for $q \leq m - n - 1$. The resulting expression for $Z_q(s)$, which holds for $|s| \geq M$, is

$$Z_q(s) = j^q \sum_{\mu=m-n}^{\infty} \frac{K_{\mu}}{(\mu-1)(\mu-2)\cdots(\mu-q)} \cdot \frac{1}{s^{\mu-q}}. \quad (8)$$

Thus the first term of the inverse power series expansion of $(-j)^{m-n-1} Z_{m-n-1}(s)$ is

$$\frac{K_{m-n}}{(m-n-1)!s},$$

where $K_{m-n} = K$. Moreover for s real and sufficiently large, this first term becomes the dominant term. Therefore K must be positive. This completes the proof.

A property of the functions $Z_q(s)$ which will be used subsequently is the fact that their real and imaginary parts are either odd or even. This is stated by Lemma 2. To prove this however, Lemma 1 will be needed.

Lemma 1. Let $h(\omega)$ be even (or odd) and integrable for $-\infty < \omega < +\infty$; let $q(\omega) = \int_{-\infty}^{\omega} h(u) du$; and let $q(\infty) = 0$. Then $q(\omega)$ is odd (or even).

Proof.

$$q(\infty) = \int_{-\infty}^{\infty} h(u) du + \int_{\infty}^{\infty} h(u) du = 0.$$

Therefore,

$$q(\omega) = \int_{-\infty}^{\omega} h(u) du = \int_{\omega}^{\infty} h(u) du.$$

Replacing u in the last integral by $-v$,

$$q(\omega) = - \int_{-\infty}^{-\omega} h(-v) dv.$$

But for $h(\omega)$ even, $h(-v)$ equals $h(v)$ so that $q(\omega)$ equals $-q(-\omega)$. (For $h(\omega)$ odd, $h(-v)$ equals $-h(v)$ so that $q(\omega)$ equals $q(-\omega)$.)

Lemma 2. For q even and less than $m - n$, the real and imaginary parts of $Z_q(s)$ are even and odd, respectively; for q odd and less than $m - n$, the real and imaginary parts of $Z_q(s)$ are odd and even, respectively.

Proof. Consider a closed path of integration in the right half s plane for the integral $\oint Z_{q-1}(s) ds$ which is composed of a straight line segment parallel to the imaginary axis and to the right of it by the distance $c \geq 0$ and a segment of a circle C_1 whose center is at the origin.

$$\oint Z_{q-1}(s) ds = \int_{c-i\infty}^{c+i\infty} Z_{q-1}(s) ds + \int_{C_1} Z_{q-1}(s) ds.$$

This integral is zero since $Z_{q-1}(s)$ is analytic for $\sigma \geq 0$. Moreover as the distance from the origin to the circular segment increases without limit, the integral along this circular segment will vanish so long as $q \leq m - n - 1$, for then $Z_{q-1}(s)$ is $o(1/s)$ as $s \rightarrow \infty$. Thus for $c \geq 0$,

$$\int_{c-i\infty}^{c+i\infty} Z_{q-1}(s) ds = 0.$$

Hence the real and imaginary parts of $Z(s)$ satisfy the hypothesis of Lemma 1. Therefore the real and imaginary parts of $Z_1(s)$ are odd and even, respectively. Furthermore they also satisfy the hypothesis of Lemma 1 so long as $m - n$ is greater than 2. This application of Lemma 1 may be made $m - n - 1$ times to obtain Lemma 2.

While the first theorem of Part II applies to all class k functions, the other theorems and corollaries hold only for certain class k functions. In particular, Theorems 10 and 11 have been proved only for those functions in class $m - n$ whose real or imaginary parts, given by the left hand side of either (6) or (7), are not only positive but are, moreover, monotonic decreasing for positive ω . This condition is stated by the following expressions where $k = m - n$. When $k = 2\nu + 1$ ($\nu = 0, 1, 2, \dots$),

$$(-1)^{\nu-1} R_{k-2}(\omega) \geq 0 \quad \text{for } \omega \geq 0 \tag{9}$$

and, when $k = 2\nu$ ($\nu = 1, 2, \dots$),

$$(-1)^\nu I_{k-2}(\omega) \geq 0 \quad \text{for } \omega \geq 0. \tag{10}$$

Actually, any system function of the form of Eq. (1) which satisfies one of these inequalities will be automatically a class k function. For the function given by the left hand side of (9) or (10) is an odd function so that one more integration according to (3) or (4) will yield one of the inequalities of (6) or (7).

Finally, it should be noted that a slightly different definition of the class k functions leads to a simpler form for Theorem 1. If the factor $(-j)^q$ in the right hand side of Eq. (2) is replaced by $(-1)^q$, the class k functions may be defined by the condition that the real part of $Z_{k-1}(j\omega)$ is non-negative for all ω . It may then be shown in this case that $Z_{k-1}(s)$ is a positive real function. However, the proofs of some of the theorems are simpler if the former definition is used.

The subclass k functions. The question of determining whether a particular $Z(s)$ is a member of any class without performing the required integrations remains. Several tests have been devised which are applicable to certain functions in a given class but not to all. These functions comprise the subclass k .

Definition. $Z(s)$ will be called a subclass k function, where $k = m - n$, if, for k odd, $R(\omega)$ has $k - 1$ changes of sign for $-\infty < \omega < +\infty$ and $R(0)$ is positive and, for k even, $I(\omega)$ has $k - 1$ changes of sign for $-\infty < \omega < \infty$ and $dI/d\omega$ at $\omega = 0$ is negative.

If a system function satisfies the conditions of the second definition then it will also satisfy those of the first definition and one of the inequalities of (9) or (10). Therefore, all the results of part II hold for all subclass k functions.

Theorem 2. All members of subclass k are members of class k . Moreover all subclass k functions satisfy one of the inequalities given by (9) or (10).

Proof. First consider the case where $m - n = 2\nu + 1$ ($\nu = 0, 1, 2, \dots$). The function $R_q(\omega)$ must have a smaller number of changes of sign in the interval $-\infty < \omega \leq \omega_i$, where ω_i is any zero of $R_{q-1}(\omega)$, than does $R_{q-1}(\omega)$. Moreover, invoking Lemma 2, it may be seen that $R_q(\omega)$ is odd or even when $R_{q-1}(\omega)$ is even or odd, respectively, so that $R_q(\omega)$ has less changes of sign in the interval $\omega_i \leq \omega < \infty$ than does $R_{q-1}(\omega)$. Thus each integration removes at least one change of sign from the finite ω axis. Moreover $R(\omega)$ has $m - n - 1$ changes of sign and $R_{m-n-1}(\infty)$ equals zero, so that each integration must remove only one change of sign. (Otherwise $R_{m-n-1}(\infty)$ would not equal zero.) Therefore $R_{m-n-2}(\omega)$ has only one change of sign which is at the origin and $R_{m-n-1}(\omega)$ has none. Since $R(0)$ is positive, $(-1)^{\nu-1} R_{m-n-2}(\omega) \geq 0$ for $0 \leq \omega < \infty$ and $(-1)^\nu R_{m-n-1}(\omega) > 0$ for $-\infty < \omega < +\infty$. This proves the theorem when $m - n$ is odd.

The same argument given in the preceding paragraph may be applied, when $m - n = 2\nu$ ($\nu = 1, 2, 3, \dots$), to obtain the remaining portion of this theorem. In this case, the fact that dI/dt at $t = 0$ is a negative quantity implies that $(-1)^\nu I_{m-n-2}(\omega) \geq 0$ for $0 \leq \omega < \infty$ and that $(-1)^{\nu+1} I_{2\nu-1}(\omega) > 0$ for all finite ω . This completes the proof.

The argument given in this proof yields a lower bound on the number of sign changes in the real or imaginary parts of a system function for real frequencies even though it may not be a class k function. This fact will be needed subsequently.

Lemma 3. Any system function having the form of Eq. (1) must have at least $m - n - 1$ changes of sign for $R(\omega)$ if $m - n$ is odd and for $I(\omega)$ if $m - n$ is even.

Proof. Otherwise $R_{m-n-1}(\infty)$ and $I_{m-n-1}(\infty)$ would not be equal to zero as they must. A system function having a positive constant multiplier and no zeros in the finite

complex frequency plane (that is, a function which is a reciprocal of a Hurwitz polynomial) will automatically be a subclass k function as stated by the next theorem. An example of a network whose transfer function is of this type is the RC ladder network which is considered by a number of authors [6-9]. The same form of transfer function holds for more generalized ladder networks [10].

Theorem 3. Let the system function $Z(s)$ have the following form,

$$Z(s) = \frac{K}{s^m + b_{m-1}s^{m-1} + \cdots + b_0} = \frac{K}{D(s)},$$

where $D(s)$ is a Hurwitz polynomial and K is positive. Then $Z(s)$ is a subclass m function.

Proof. The proof depends upon a known property of Hurwitz polynomials, namely, that all the zeros of the even and odd parts of a Hurwitz polynomial of m th degree are simple and purely imaginary [11]. Thus, if m is odd, $R(\omega)$ has $m - 1$ real, simple zeros and, if m is even, $I(\omega)$ has $m - 1$ real, simple zeros. Furthermore, the series expansion of $Z(s)$ is

$$Z(s) = \frac{K}{b_0} - \frac{Kb_1}{b_0} s + \cdots.$$

Since $D(s)$ is a Hurwitz polynomial, all its coefficients are positive [12]. Therefore,

$$R(0) = \frac{K}{b_0} > 0$$

$$\left. \frac{dI}{d\omega} \right|_{\omega=0} = -\frac{Kb_1}{b_0} < 0.$$

Hence, all the conditions of the definition of a subclass m function are satisfied.

Tests for a subclass k function. Two tests have been devised which may be used to determine whether a given system function is a subclass k function. These tests determine the number of real zeros in the real or imaginary parts of the system function whose expression for real frequencies is given by (11).

$$Z(j\omega) = K \frac{N(j\omega)D(-j\omega)}{|D(j\omega)|^2}. \quad (11)$$

Since all the roots of $D(s)$ have negative real parts, $|D(j\omega)|^2$ is positive and finite for all finite ω . Thus, the zeros of the real and imaginary parts of $Z(j\omega)$ are the zeros of the real and imaginary parts of $N(j\omega) D(-j\omega)$. Let $P(\omega^2)$ be the even part of $N(j\omega) D(-j\omega)$ and let $\omega Q(\omega^2)$ be its odd part. Replacing the variable ω^2 by x , the following test may be stated which is based on Descartes' rule of signs [13].

Theorem 4. A system function $Z(s)$ is a subclass $m - n$ function if the following conditions hold. For $m - n$ odd, the number of sign changes in the coefficients of $P(x)$ is $(m - n - 1)/2$ and $P(0)$ is positive. For $m - n$ even, the number of sign changes in the coefficients of $Q(x)$ is $(m - n - 2)/2$ and $Q(0)$ is negative.

Proof. First consider the case where $m - n$ is odd. The number of real roots of $R(\omega)$ equals twice the number of positive roots of $P(x)$. By Descartes' rule of signs [13] the number of positive roots of $P(x)$ is less than or equal to the number of variations of

sign in the coefficients of $P(x)$. So by hypothesis, the number of real roots of $R(\omega)$ is less than or equal to $m - n - 1$. Moreover, by Lemma 3, the number of changes of sign for $R(\omega)$, which is less than or equal to the number of real roots of $R(\omega)$, is at least $m - n - 1$. Thus $W(\omega)$ has exactly $m - n - 1$ real simple roots and the conditions for the definition of a subclass $m - n$ function are fulfilled.

The proof for the case where $m - n$ is even is the same except that now the number of real roots of $I(\omega)$ equals twice the number of positive roots of $Q(x)$ plus one more.

This theorem may be applied to determine another general type of subclass k function having only one zero as given by Eq. (12).

$$Z(s) = K \frac{s + a_0}{s^m + b_{m-1}s^{m-1} + \cdots + b_0} = K \frac{s + a_0}{D(s)}. \quad (12)$$

A condition that the coefficients of Hurwitz polynomials satisfy is

$$b_{m-1} > \frac{b_{m-3}}{b_{m-2}} > \frac{b_{m-5}}{b_{m-4}} > \cdots > 0. \quad (13)$$

The next to the last term in (13) is b_1/b_2 when m is even and b_0/b_1 when m is odd. (In those cases where $m = 3$ or $m = 4$, these inequalities are a consequence of the Hurwitz criterion [14]. Professor C. F. Rehberg has proven that they hold for Hurwitz polynomials of any degree so that the conditions (13) may be omitted from the hypothesis of the following theorem.)

Theorem 5. If a system function $Z(s)$, having one zero, has the form of Eq. (12), if its coefficients satisfy the inequalities (13) and if the position of the zero satisfies $b_{m-1} \geq a_0 > b_0/b_1$ for m odd and $b_{m-1} \geq a_0 > 0$ for m even, then $Z(s)$ is a subclass $m - 1$ function.

Proof. It will be shown that a system function which satisfies this hypothesis will also satisfy the hypothesis of Theorem 4. First consider the case where m is odd. Since $m - 1$ is even, $Q(x)$ is to be calculated and the number of changes of sign in its coefficients determined.

$$\begin{aligned} Q(x) = j^{m-1} \Big[& (b_{m-1} - a_0)x^{(m-1)/2} + (a_0 b_{m-2} - b_{m-3})x^{(m-3)/2} \\ & + (b_{m-5} - a_0 b_{m-4})x^{(m-5)/2} + \cdots + \frac{b_0 - a_0 b_1}{j^{m-1}} \Big]. \end{aligned}$$

Now it is evident that if the conditions (13) are satisfied and if $b_{m-1} \geq a_0 > b_0/b_1$ then the coefficients of $Q(x)$ have $(m - 3)/2$ changes in sign. But this number equals $(m - n - 2)/2$ for n equal to 1. Furthermore, $Q(0)$ is negative for $a_0 > b_0/b_1$. Thus the hypothesis of Theorem 4 is satisfied.

For the case where m is even, $P(x)$ may be found to be

$$\begin{aligned} P(x) = j^m \Big[& (a_0 - b_{m-1})x^{m/2} + (b_{m-3} - a_0 b_{m-2})x^{(m-2)/2} \\ & + (a_0 b_{m-4} - b_{m-5})x^{(m-4)/2} + \cdots + \frac{a_0 b_0}{j^m} \Big]. \end{aligned}$$

Again, if the conditions (13) are satisfied and if $b_{m-1} \geq a_0 > 0$ then the coefficients of $P(x)$ have $(m - 2)/2$ changes in sign which is $(m - n - 1)/2$ in number for n equal to 1. Furthermore, since $D(s)$ is a Hurwitz polynomial, b_0 is positive. Therefore, $P(0)$ is positive for $a_0 > 0$ and the hypothesis of Theorem 4 is again satisfied.

Descartes' rule of signs only yields an upper bound on the number of positive roots of a polynomial. Sturm's theorem [13] which determines exactly the number of positive roots of a polynomial (a multiple root being counted as a single root) is the basis of the following test.

Theorem 6. Let $f_0(x)$ equal $P(x)$ for $m - n$ odd and equal $Q(x)$ for $m - n$ even. Let $f_1(x)$ equal df_0/dx . Let $f_2(x)$ be the negative of the remainder obtained by dividing $f_0(x)$ by $f_1(x)$. Let $f_3(x)$ be the negative of the remainder obtained by dividing $f_1(x)$ by $f_2(x)$. Let this process be continued as indicated below until the last remainder $f_k(x)$ is a constant or a polynomial which never changes sign.

$$\begin{aligned} f_0(x) &= q_1(x)f_1(x) - f_2(x) \\ f_1(x) &= q_2(x)f_2(x) - f_3(x) \\ \dots & \\ f_{k-2}(x) &= q_{k-1}(x)f_{k-1}(x) - f_k(x). \end{aligned}$$

Finally, let V_0 be the number of sign changes in the sequence $f_0(0), f_1(0), \dots, f_k(0)$ and let V_∞ be the number of sign changes in the sequence $f_0(\infty), f_1(\infty), \dots, f_k(\infty)$. If, for $m - n$ odd, $V_0 - V_\infty$ equals $(m - n - 1)/2$ and $P(0)$ is positive or, for $m - n$ even, $V_0 - V_\infty$ equals $(m - n - 2)/2$ and $Q(0)$ is negative, then $Z(s)$ is a subclass $m - n$ function.

Proof. The hypothesis of this theorem is simply a statement of Sturm's test. For instance, when $m - n$ is odd, the excess of V_0 over V_∞ is exactly the number of real positive roots of $P(x)$. When this equals $(m - n - 1)/2$, $R(\omega)$ will have $m - n - 1$ real zeros and the conditions for a subclass $m - n$ function will be satisfied. A similar situation exists when $m - n$ is even.

The phase functions of subclass k functions. The phase angle of a subclass k function for real frequencies has several properties which will be developed in this section. This phase function $\varphi(\omega)$ is defined by

$$Z(j\omega) = |Z(j\omega)| \exp [j\varphi(\omega)]. \quad (14)$$

Of course, any multiple of 2π may be added to $\varphi(\omega)$ without affecting the value of $Z(j\omega)$. In order to deal with a unique phase function, the following convention will be adopted. Consider the factored form of the system function given by (15) where the poles are denoted by the symbols ρ_i and the zeros by the symbols μ_i .

$$Z(s) = K \frac{(s - \mu_1)(s - \mu_2) \cdots (s - \mu_n)}{(s - \rho_1)(s - \rho_2) \cdots (s - \rho_m)}. \quad (15)$$

It will be presumed that at any real frequency the phase angle of any factor in this expression, which is determined by the angle of the vector extending from the pole or zero to the particular frequency in question on the imaginary axis, remains within the bounds $3\pi/2$ and $-\pi/2$. Thus as ω increases from $-\infty$ to $+\infty$, the phase angle for a factor of a pole or zero in the left half s plane will increase from $-\pi/2$ to $\pi/2$ whereas this angle for a pole or zero in the right half s plane will decrease from $3\pi/2$ to $\pi/2$. When the pole or zero occurs on the imaginary axis, the phase angle for its factor will be $-\pi/2$ when ω is smaller than the pole or zero and $\pi/2$ when ω is larger. Such angles are illustrated in Fig. 1 where the symbol ψ_i denotes the phase angle for the factor of a zero and the symbol θ_i denotes this angle for a pole. As was stated previously in Theorem

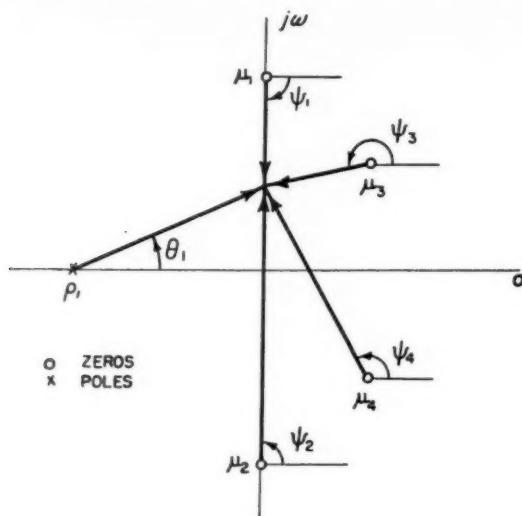


FIG. 1. Illustration of the Convention for Measuring Phase Angles.

1 the constant multiplier K must be a positive number if $Z(s)$ is a class k function. Thus by this convention the phase angle for $Z(j\omega)$ is unique and it is given by

$$\varphi(\omega) = \sum_{i=1}^n \psi_i - \sum_{i=1}^m \theta_i . \quad (16)$$

Another test for the subclass k functions may be constructed using the phase function $\varphi(\omega)$. It also determines the number of real zeros that $R(\omega)$ and $I(\omega)$ possess.

Theorem 7. If the system function $Z(s)$ is given by Eq. (1), if its phase function at real frequencies $\varphi(\omega)$ is continuous, if $d\varphi/d\omega < 0$, and if $-(m-n)\pi/2 < \varphi(\omega) < (m-n)\pi/2$ for $-\infty < \omega < \infty$, then $Z(s)$ is a subclass $m-n$ function.

Proof. Whenever the phase function $\varphi(\omega)$ equals an odd multiple of $\pi/2$, the real part of the system function at real frequencies $R(\omega)$ equals zero and, whenever $\varphi(\omega)$ equals zero or a multiple of π , the imaginary part $I(\omega)$ equals zero. Thus, under the conditions of the hypothesis, as ω increases continuously from $-\infty$ to $+\infty$, $\varphi(\omega)$ will decrease continuously from $+(m-n)\pi/2$ to $-(m-n)\pi/2$ and the number of times $R(\omega)$ or $I(\omega)$ changes sign may be determined by counting the number of times $\varphi(\omega)$ passes through an odd multiple of $\pi/2$ or a multiple of π . When $m-n$ is odd, $R(\omega)$ has $m-n-1$ changes of sign and when $m-n$ is even, $I(\omega)$ has $m-n-1$ changes of sign.

Furthermore, the denominator of $Z(s)$ is a Hurwitz polynomial so that its magnitude at finite real frequencies is always finite and positive. Moreover $Z(s)$ cannot have any zeros on the imaginary axis for otherwise the phase function $\varphi(\omega)$ would have a discontinuity at each such zero. Thus the magnitude of $Z(j\omega)$ is always finite and non-zero for finite ω . Since $\varphi(0)$ equals zero ($\varphi(\omega)$ being an odd function), $R(0)$ is positive. The fact that $d\varphi/d\omega < 0$ at $\omega = 0$ implies that $dI/d\omega$ is negative at $\omega = 0$.

Therefore, all the conditions for the definition of a subclass $m - n$ function are satisfied.

The hypothesis of this theorem is really much too strong. The condition that $\varphi(\omega)$ is strictly monotonic decreasing may be replaced by the following conditions which encompass a much larger set of functions. For $m - n$ odd, $\varphi(\omega)$ equals an odd multiple of $\pi/2$ exactly $m - n - 1$ times and, for $m - n$ even, $\varphi(\omega)$ equals a multiple of π or zero exactly $m - n - 1$ times; moreover $d\varphi/d\omega < 0$ at $\omega = 0$. In this case the proof is practically the same.

It can easily be seen that this theorem may be used in an alternate proof of Theorem 3. Since all the poles of the system function which is the reciprocal of a Hurwitz polynomial are in the left half plane and since there are no zeros, the phase function $\varphi(\omega)$ is strictly monotonic decreasing and satisfies the hypothesis of Theorem 7.

Several properties held by the phase function at real frequencies $\varphi(\omega)$ of a subclass k function are stated by the following theorem.

Theorem 8. If the system function $Z(s)$, given by Eq. (1), is a subclass $m - n$ function, then it satisfies the following conditions.

- i. $Z(s)$ has no zeros in the right half s plane; that is, $Z(s)$ is a minimum phase function.
- ii. $Z(s)$ has no zeros on the imaginary axis; that is, $\varphi(\omega)$ is a continuous function.
- iii. $-(m - n)\pi/2 < \varphi(\omega) < (m - n)\pi/2$ for $-\infty < \omega < \infty$.

Proof. i and ii. It is convenient to prove the first two statements of the conclusion together. Assume that there are p zeros in the right half s plane and q zeros on the imaginary axis so that the number of zeros in the left half s plane is $(n - p - q)$. All the poles are in the left half s plane. As ω increases from $-\infty$ to $+\infty$, the angle θ_i corresponding to the pole ρ_i will increase continuously from $-\pi/2$ to $\pi/2$. The angles ψ_i for any zeros in the left half s plane will behave similarly. However these angles for zeros in the right half s plane will decrease from $3\pi/2$ to $\pi/2$. Finally a zero on the imaginary axis will have an angle for its factor which is $\pm\pi/2$ at all ω except at the zero where there will be a discontinuity of magnitude π .

Now consider the phase functions $\varphi'(\omega)$ determined by all the poles and only those zeros which are off the imaginary axis. This function is continuous.

$$\varphi'(\omega) = \sum_{i=1}^{n-q} \psi_i - \sum_{i=1}^m \theta_i .$$

The number of times $R(\omega)$ or $I(\omega)$ changes sign must be at least as great as the number of times the function $\varphi'(\omega)$ varies continuously and monotonically through odd multiples of $\pi/2$ or multiples of π , respectively. For, the contribution to the phase function $\varphi(\omega)$ due to the zeros on the imaginary axis is a step function each of whose discontinuities is a multiple of π and such a contribution will only yield additional zeros to $R(\omega)$ and $I(\omega)$. The function $\varphi'(\omega)$ varies continuously over a range at least as large as $[m - (n - p - q) + p]\pi$. The only way for $R(\omega)$ or $I(\omega)$ to have only $m - n - 1$ changes of sign is for this range to be no greater than $(m - n)\pi$. Thus both p and q must be zero.

iii. By the above argument, all the poles and zeros of $Z(s)$ are in the left half s plane so that $\varphi(0)$ equals zero. Moreover it has been shown that the range of variation for $\varphi(\omega)$ can be no greater than $(m - n)\pi$. Thus $\varphi(\omega)$ is an odd function and it is bounded by $\pm(m - n)\pi/2$ for $-\infty < \omega < +\infty$.

Part II. Bounds on the impulse and step responses. As is well known, the unit impulse response $W(t)$ is related to the system function by the Fourier transform

$$W(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} Z(j\omega) \exp(j\omega t) d\omega. \quad (17)$$

It will be presumed that all input functions are applied at $t = 0$ so that the response for any physically realizable system must be zero for negative values of time. This result may also be derived from the condition that $Z(s)$ has no poles in the right half s plane. Because of this, $W(t)$ may be represented either by Eqs. (18) or (19) for positive values of time.

$$W(t) = \frac{2}{\pi} \int_0^{\infty} R(\omega) \cos \omega t d\omega, \quad t \geq 0 \quad (18)$$

$$W(t) = -\frac{2}{\pi} \int_0^{\infty} I(\omega) \sin \omega t d\omega, \quad t > 0 \quad (19)$$

It has been shown previously that when the frequency responses of networks are restricted in various ways the transient responses are bounded [1-3]. Specifically, when the real part of a system function is of constant sign and the system function has one more pole than the number of its zeros, the magnitude of the corresponding unit impulse response is bounded by the constant multiplier K of the system function. Similarly when the imaginary part of a system function with two more poles than zeros is of constant sign for positive frequencies, then the magnitude of the unit impulse response is bounded by Kt . The restriction on the frequency characteristic of a class k system function leads to a generalization of these two conclusions. This generalization, which is given by the following theorem, states that the magnitude of the unit impulse response is bounded by $Kt^{m-n-1}/(m-n-1)!$, the symbols being defined in Eq. (1). (By the initial value theorem, the upper bound is seen to be the initial value of the unit impulse response.) It follows that the magnitude of the unit step response will be bounded by $Kt^{m-n}/(m-n)!$ so that the rise time from zero to one will always be greater than $[(m-n)!r/K]^{1/(m-n)}$ where r equals Ka_0/b_0 (presuming that $a_0 > 0$).

Theorem 9. If the system function $Z(s)$, given by Eq. (1), is a class $m - n$ function, then the corresponding unit impulse response $W(t)$ is bounded by the following expression for $t \geq 0$.

$$|W(t)| \leq \frac{Kt^{m-n-1}}{(m-n-1)!}. \quad (20)$$

Proof. For $m - n = 2\nu + 1$ ($\nu = 0, 1, 2, \dots$), upon integrating by parts expression (18) 2ν times and using the fact that $R_q(\infty) = 0$ for $q < m - n$, the following may be obtained

$$W(t) = (-1)^{\nu} t^{2\nu} \frac{2}{\pi} \int_0^{\infty} R_{2\nu}(\omega) \cos \omega t d\omega. \quad (21)$$

But by definition of a class $m - n$ function $(-1)^{\nu} R_{2\nu}(\omega) \geq 0$ for all ω and so

$$|W(t)| \leq (-1)^{\nu} t^{2\nu} \frac{2}{\pi} \int_0^{\infty} R_{2\nu}(\omega) d\omega.$$

By integrating $Z_{2r}(s)$ around the $\sigma > 0$ half plane, the integral, $\int_0^\infty R_{2r}(\omega) d\omega$, is found to equal $(-1)^r K\pi/2(2r)!$. Thus the conclusion of this theorem is obtained for the case where $m - n$ is odd.

For $m - n = 2r$ ($r = 1, 2, 3, \dots$), a similar argument may be applied to (19). Integration by parts $2r - 1$ times yields

$$W(t) = (-1)^{r+1} t^{2r-1} \frac{2}{\pi} \int_0^\infty I_{2r-1}(\omega) \cos \omega t d\omega. \quad (22)$$

Again by a definition of a class $m - n$ function, $(-1)^{r+1} I_{2r-1}(\omega) \geq 0$ for all ω so that

$$|W(t)| \leq (-1)^{r+1} t^{2r-1} \frac{2}{\pi} \int_0^\infty I_{2r-1}(\omega) d\omega.$$

Integrating $Z_{2r-1}(s)$ around the $\sigma > 0$ half plane, the integral on the right hand side of the last expression is found to be equal to $(-1)^{r+1} K\pi/2(2r - 1)!$. This yields the conclusion once more.

The next two theorems depend upon two inequalities that the sine function satisfies and which were proved in a previous paper [3].

Lemma 4. For $0 \leq y < 1$, $x \geq 0$ and N a positive integer,

$$\sin x \leq Q_0 x + \sum_{p=1}^N \frac{yQ_{2p}}{p} \sin \frac{px}{y} \quad (23)$$

and

$$\sin x \geq -Q_0 x + \sum_{p=1}^N (-1)^{p+1} \frac{yQ_{2p}}{p} \sin \frac{px}{y}, \quad (24)$$

where

$$Q_0 = 1 + \sum_{k=1}^N \frac{(-y^2)(1^2 - y^2) \cdots [(k-1)^2 - y^2]}{(k!)^2}, \quad (25)$$

$$Q_{2p} = (-1)^p 2 \sum_{k=p}^N \frac{(-y^2)(1^2 - y^2) \cdots [(k-1)^2 - y^2]}{(k-p)!(k+p)!}, \quad p = 1, 2, \dots, N. \quad (26)$$

Any system function which satisfies the inequality (9) or the inequality (10) (as has been noted before, all subclass k functions are of this type) will have the property that, when its corresponding impulse response is bounded beyond a certain time, then other bounds on this response will be determined before this time. The physical significance of this is that the more rapidly an impulse response "settles down" the less violent must this response be. This result is given by the next theorem which is a generalization of Theorems 1 and 2 in [15]. More precisely, the conclusions of [15] are special cases of the following obtained by setting $m - n$ equal to one or two and making some trivial changes in the notation and normalization.

Theorem 10. Let the system function $Z(s)$, given by Eq. (1), satisfy the inequality (9) if $m - n$ is odd or the inequality (10) if $m - n$ is even. If the magnitude of the corresponding unit impulse response $W(t)$ is less than or equal to M for $t \geq \tau$ where M is a positive number, then for $0 \leq y < 1$,

$$|W(y\tau)| \leq \frac{K \sin \pi y}{(m-n-1)! \pi y} (y\tau)^{m-n-1} + \frac{2My^{m-n} \sin \pi y}{\pi} \sum_{v=1}^{\infty} \frac{1}{v^{m-n-1} (v^2 - y^2)}. \quad (27)$$

Proof. First consider the case where $m - n = 2\nu + 1$ ($\nu = 0, 1, 2, \dots$). Integrating (18) by parts $2\nu - 1$ times, (28) is obtained.

$$W(t) = (-1)^{\nu-1} t^{2\nu-1} \frac{2}{\pi} \int_0^\infty R_{2\nu-1}(\omega) \sin \omega t \, d\omega. \quad (28)$$

By hypothesis, $(-1)^{\nu-1} R_{2\nu-1}(\omega)$ is non-negative. Therefore, setting x equal to ωt , the sine function may be replaced by the right hand side of (23) and the result integrated term by term to yield

$$W(t) \leq (-1)^{\nu-1} t^{2\nu} Q_0 \frac{2}{\pi} \int_0^\infty \omega R_{2\nu-1}(\omega) \, d\omega + \sum_{p=1}^N \left(\frac{y}{p}\right)^{2\nu} Q_{2p} W\left(\frac{pt}{y}\right).$$

Moreover,

$$\int_0^\infty \omega R_{2\nu-1}(\omega) \, d\omega = - \int_0^\infty R_{2\nu}(\omega) \, d\omega.$$

The value for this last integral may be found by integrating $Z_{2\nu}(s)$ around the $\sigma > 0$ half plane

$$-\int_0^\infty R_{2\nu}(\omega) \, d\omega = \frac{(-1)^{\nu+1} \pi K}{2(2\nu)!}.$$

Therefore,

$$W(t) \leq \frac{K Q_0 t^{2\nu}}{(2\nu)!} + \sum_{p=1}^N \left(\frac{y}{p}\right)^{2\nu} Q_{2p} W\left(\frac{pt}{y}\right). \quad (29)$$

Furthermore, the $W(t)$ are bounded for all t since

$$|W(t)| \leq \frac{2}{\pi} \int_0^\infty |R(\omega)| \, d\omega < \infty.$$

Thus it may be found that the double series obtained by letting N go to infinity in the right hand side of (29) converges absolutely for $0 \leq y < 1$. So summing over the k in Eqs. (25) and (26), the following may be obtained where the Q_{2p} converge to the q_{2p} as shown in [16].

$$W(t) \leq \frac{K q_0 t^{2\nu}}{(2\nu)!} + \sum_{p=1}^\infty \left(\frac{y}{p}\right)^{2\nu} q_{2p} W\left(\frac{pt}{y}\right), \quad (30)$$

where

$$q_0 = \frac{\sin \pi y}{\pi y}$$

$$q_{2p} = (-1)^{p+1} \frac{2y \sin \pi y}{\pi(p^2 - y^2)}.$$

Now if t/y is set equal to τ and $W(pt/y)$ is replaced by M for p odd and by $-M$ for p even, the upper bound of (27) is achieved. The lower bound is similarly obtained by use of (24) rather than (23).

The conclusion of this theorem may also be achieved in the case where $m - n = 2\nu$ ($\nu = 1, 2, 3, \dots$) by integrating (19) by parts $2\nu - 2$ times and then proceeding in the same way.

Finally bounds which are similar to those of Theorem 10 exist on the step responses of those system functions which satisfy either inequality (9) or (10) and whose constant a_0 is greater than zero. Once more, it may be noted that all subclass k functions are of this type. The unit step response $A(t)$ is related to the unit impulse response $W(t)$ by the expression,

$$A(t) = \int_0^t W(\xi) d\xi. \quad (31)$$

The following theorem is once again a generalization of Theorem 3 of [15].

Theorem 11. Let the system function $Z(s)$, given by Eq. (1), satisfy the inequality (9) if $m - n$ is odd or the inequality (10) if $m - n$ is even and let a_0 be positive. If the corresponding unit step response $A(t)$ is bounded by $(1 \pm \gamma)r$ for $t \geq \tau$ where $r = K a_0 / b_0 > 0$ and γ is a positive number, then, for $0 \leq y < 1$,

$$A(y\tau) \leq \frac{y^{m-n-1} \sin \pi y}{\pi} \left\{ \frac{K\tau^{m-n}}{(m-n)!} + 2y^2 r \left[\sum_{\nu=1}^{\infty} \frac{(-1)^{\nu+1}}{\nu^{m-n} (\nu^2 - y^2)} + \gamma \sum_{\nu=1}^{\infty} \frac{1}{\nu^{m-n} (\nu^2 - y^2)} \right] \right\} \quad (32)$$

and

$$A(y\tau) \geq -\frac{y^{m-n-1} \sin \pi y}{\pi} \left\{ \frac{K\tau^{m-n}}{(m-n)!} - 2y^2 r(1-\gamma) \sum_{\nu=1}^{\infty} \frac{1}{\nu^{m-n} (\nu^2 - y^2)} \right\}. \quad (33)$$

Proof. To obtain the upper bound (32) for both cases where $m - n$ is odd and $m - n$ is even, the proof proceeds in the same way as in the preceding theorem until the inequality given by (29) is obtained where 2ν is replaced by $m - n - 1$. Integrating according to (31),

$$A(t) \leq \frac{KQ_0 t^{m-n}}{(m-n)!} + \sum_{p=1}^N \left(\frac{y}{p} \right)^{m-n} Q_{2p} A \left(\frac{pt}{y} \right). \quad (34)$$

However,

$$|A(t)| \leq \frac{2t}{\pi} \int_0^\infty |R(\omega)| d\omega \quad \text{for } t \geq 0.$$

Thus $A(t)$ is bounded for $0 \leq t < \infty$ and so it may be seen that the double series obtained by letting N go to infinity in the right hand side of (34) converges absolutely. Upon letting N go to infinity and summing over the k in Eqs. (25) and (26), the following may be obtained where the q_{2p} are given below expression (30)

$$A(t) \leq \frac{Kq_0 t^{m-n}}{(m-n)!} + \sum_{p=1}^N \left(\frac{y}{p} \right)^{m-n} q_{2p} A \left(\frac{pt}{y} \right). \quad (35)$$

Now setting t/y equal to τ and replacing $A(pt/y)$ by $(1 + \gamma)r$ for p odd and by $(1 - \gamma)r$ for p even, expression (32) is achieved.

Expression (33) may be derived in a similar fashion using (24). In this case, the expression corresponding to (34) is the following

$$A(t) \geq -\frac{KQ_0 t^{m-n}}{(m-n)!} + \sum_{p=1}^N (-1)^{p+1} \left(\frac{y}{p} \right)^{m-n} Q_{2p} A \left(\frac{pt}{y} \right).$$

As N goes to infinity the double series converges absolutely again. Also the $A(pt/y)$ must be replaced by $(1 - \gamma)r$ for all p to maintain the inequality. This completes the proof.

For $m - n$ equal to one or two, some of the infinite series given in (27), (32) and (33) may be expressed in closed form as follows.

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu^2 - y^2} = \frac{1}{2y^2} - \frac{\pi}{2y} \cot \pi y, \quad 0 \leq y < 1$$

$$2y^2 \sum_{\nu=1}^{\infty} \frac{1}{\nu(\nu^2 - y^2)} = -[\Psi(y) + \Psi(-y) + 2g], \quad 0 < y < 1.$$

Here $\Psi(y)$ is the digamma function [17] defined by

$$\Psi(y) = \frac{d}{dy} \log \Gamma(y)$$

and g is Euler's constant defined by

$$g = \lim_{n \rightarrow 0} \left(\sum_{\nu=1}^n \frac{1}{\nu} - \log n \right) = .5772 \dots$$

The results of Theorems 10 and 11 and the following corollaries 11a and 11b are not the best possible and can be improved. For it was presumed in the proofs that $W(pr)$ and $A(pr)$ equal their bounds $\pm M$ and $(1 \pm \gamma)r$ for all positive integers p . This behavior is impossible since $W(\infty)$ equals zero and $A(\infty)$ equals r .

Restrictions on the rise time and settling time. Theorem 11 yields restrictions on the shape factors of the unit step response which are again generalizations of results that have appeared previously [3]. Defining the rise time T_γ as the time it takes for the unit step response to first cross the final value line after the input has been impressed (see Fig. 2), it is found that a lower bound exists on this rise time which becomes larger

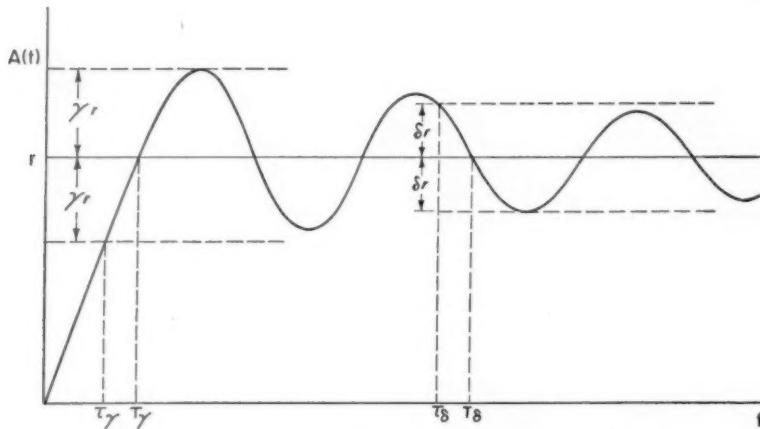


FIG. 2. Illustration of the Shape Factors for the Unit Step Response.

as the maximum overshoot or undershoot γ becomes smaller. That is, the rise time and maximum overshoot or undershoot of the unit step response define a point on the plane of Fig. 3 and this point must lie above the curve for the appropriate $m - n$. Furthermore, let T_δ be the least time at which the unit step response crosses the final value line and beyond which the overshoots and undershoots are less than or equal to δ . (Some

of the overshoots and undershoots may be greater than δ before T_i). In this case the curves of Fig. 3 still apply giving lower bounds on this shape factor T_i . These results may be stated as follows.

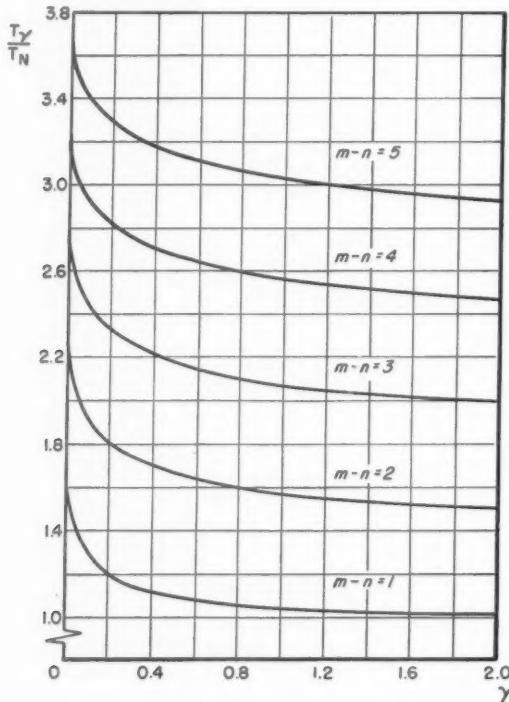


FIG. 3. The Envelopes of the Lower Bounds on the Rise Time of the Unit Step Response Given by Corollary 11a.

Corollary 11a. Let the system function $Z(s)$, given by Eq. (1), satisfy the inequality (9) if $m - n$ is odd or the inequality (10) if $m - n$ is even and let a_0 be positive. If the corresponding unit step response $A(t)$ is bounded by $(1 \pm \gamma)r$ for $t \geq T_\gamma$ where $r = K a_0/b_0$, γ is a positive number, and $A(T_\gamma) = r$, then

$$T_\gamma > \left\{ \frac{(m-n)!r}{K} \left[\frac{\pi y}{\sin \pi y} - 2y^{m-n+2} \sum_{v=1}^{\infty} \frac{(-1)^{v+1}}{v^{m-n}(v^2 - y^2)} - 2\gamma y^{m-n+2} \sum_{v=1}^{\infty} \frac{1}{v^{m-n}(v^2 - y^2)} \right] \right\}^{1/(m-n)}, \quad (36)$$

where $0 \leq y < 1$.

This corollary follows immediately from expression (32) if $y\tau$ is set equal to T_γ so that $A(y\tau)$ equals r . Since $A(\infty) = r$, equality between the two sides of (36) is impossible.

For a given y and $m - n$, the right hand side of expression (36) is a function of γ . As the parameter y is varied between zero and one, a family of curves is generated on the plane of Fig. 3 and T_γ must be greater than the envelope of this family of curves.

These envelopes for several values of $m - n$ are shown in Fig. 3, where T_N is a normalization factor given by

$$T_N = \left(\frac{r}{K} \right)^{1/(m-n)}. \quad (37)$$

As y approaches one, the envelopes approach a point on the ordinate axis given by

$$(m-n)! \left[\frac{2(m-n)+1}{2} + 2 \sum_{v=2}^{\infty} \frac{(-1)^v}{v^{m-n}(v^2-1)} \right]. \quad (38)$$

As y approaches zero, the envelopes progress infinitely to the right approaching the horizontal line whose ordinate is $(m-n)!$.

A settling time τ_s for the unit step response may be defined as the least time beyond which the unit step response remains within the bounds $(1 \pm \delta)r$. This is illustrated in Fig. 2. Again Theorem 11 implies a lower bound on τ_s as shown in Fig. 4. For a given δ ,

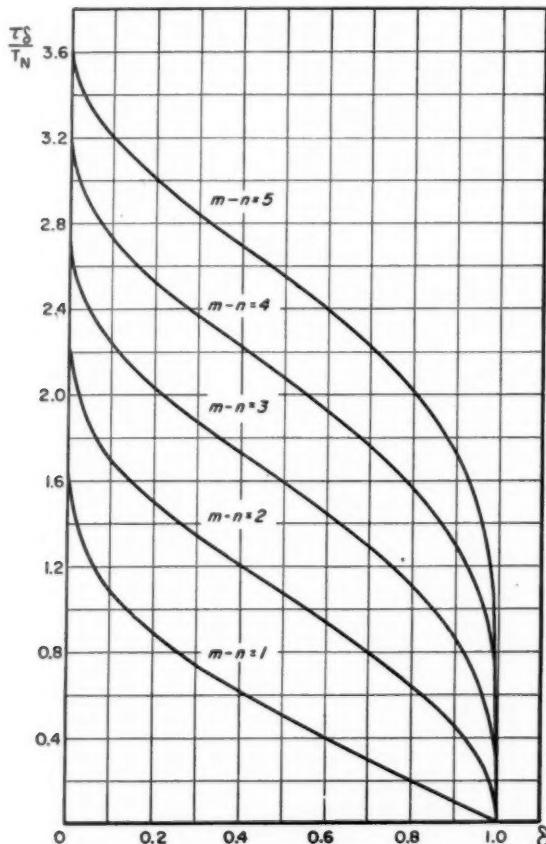


FIG. 4. The Envelopes of the Lower Bounds on the Settling Time of the Unit Step Response Given by Corollary 11b.

the settling time τ_s must be greater than the curve corresponding to the appropriate $m - n$. These curves are an immediate consequence of corollary 11b.

Corollary 11b. Let the system function $Z(s)$, given by Eq. (1), satisfy the inequality (9) if $m - n$ is odd or the inequality (10) if $m - n$ is even and let a_0 be positive. If the corresponding unit step response $A(t)$ is bounded by $(1 \pm \delta)r$ for $t \geq \tau_s$ where $r = Ka_0/b_0$ and δ is a positive number between zero and one, then

$$\tau_s > \left\{ \frac{(m-n)!r}{K} \left[\frac{\pi y}{\sin \pi y} (1 - \delta) - 2y^{m-n+2} \sum_{v=1}^{\infty} \frac{(-1)^{v+1}}{v^{m-n}(v^2 - y^2)} - 2\delta y^{m-n+2} \sum_{v=1}^{\infty} \frac{1}{v^{m-n}(v^2 - y^2)} \right] \right\}^{1/(m-n)}, \quad (39)$$

where $0 \leq y < 1$.

Setting $y\tau$ equal to τ_s , $A(y\tau)$ may be replaced by $(1 - \delta)r$, τ by τ_s/y and γ by δ in expression (32). This will yield expression (39) upon rearrangement. Again $A(\infty) = r$ so that equality between the two sides of expression (39) cannot be achieved.

Once again, a family of curves is generated on the plane of Fig. 4 by the right hand side of expression (39) when the parameter y is varied between zero and one. The settling time τ_s must be greater than the envelopes of these families of curves for various values of $m - n$. These envelopes are shown in Fig. 4. As y approaches one, the envelopes approach a point on the ordinate axis which is given by expression (38) and, as y approaches zero, they approach the value, one, on the abscissa axis.

APPENDIX I

Examples and applications. 1. As an illustration of the bounds holding on the impulse response of a system function which is the reciprocal of a Hurwitz polynomial, the following system function is considered.

$$Z(s) = \frac{1.241}{(s + .591)^2 + (.806)^2}.$$

The corresponding unit impulse response is

$$W(t) = 1.537e^{-.591t} \sin .806t$$

and this is plotted in Fig. 5. Since $Z(s)$ is the reciprocal of a Hurwitz polynomial of second degree, it is automatically a Subclass 2 function by Theorem 3. Thus all the bounds developed in Part II apply to its corresponding unit impulse and unit step responses. The bound, 1.241, on the magnitude of the unit impulse response given by Theorem 9 is readily seen to hold. Furthermore, as seen in Fig. 5, $|W(t)|$ is less than .063 for all t greater than 3.486. Thus by Theorem 10, $|W(y\tau)|$ is bounded in the interval $0 \leq t < 3.486$ and these bounds are also shown in Fig. 5. Of course other bounds on $|W(t)|$ could be obtained by choosing other possible values of M and τ .

2. An example of a Subclass 4 function is the following system function

$$Z(s) = \frac{(s + 1.25)^2 + .5625}{(s + 1)(s + 2)[(s + 1)^2 + 1][(s + 3)^2 + 9]}.$$

To show that this is indeed a Subclass 4 function, the function $Q(x)$, which is defined

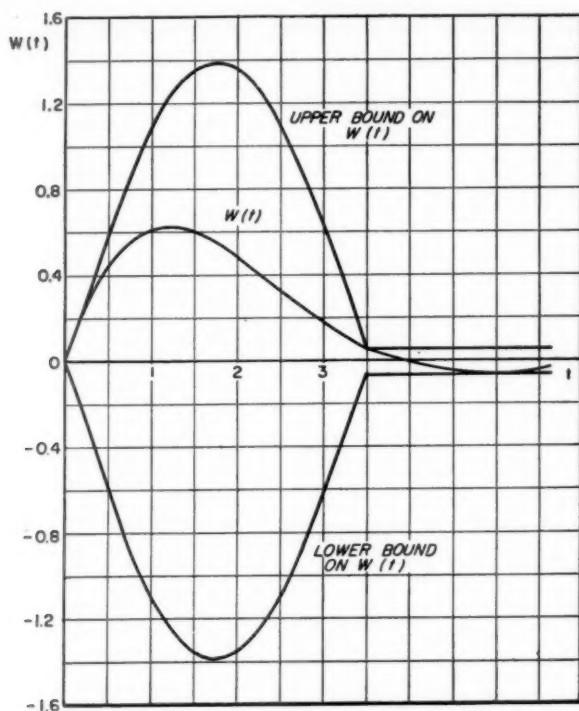


FIG. 5. Illustration of the Bounds of Theorem 10 in the Case Where $K = 1.241$, $M = 0.063$, $m - n = 2$, and $\tau = 3.486$.

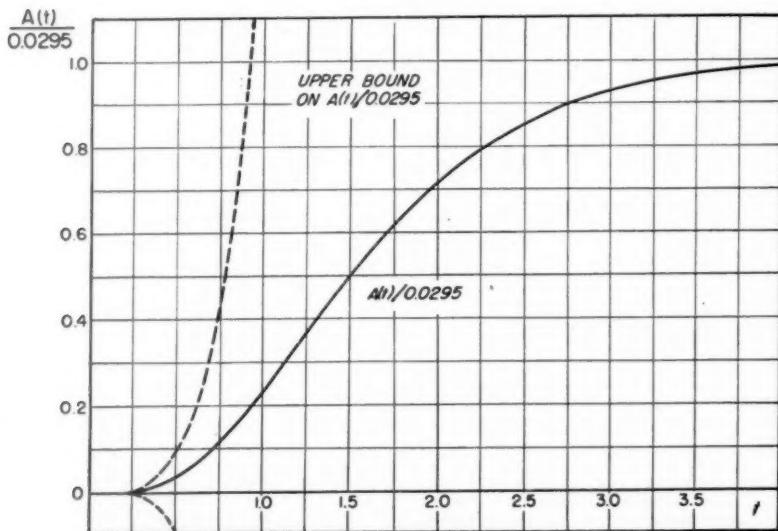


FIG. 6. Bound on the Unit Step Response Discussed in the Second Example.

immediately before the statement of Theorem 4, may be found to be

$$Q(x) = 8.500x^3 - 38.38x^2 - 66.00x - 253.5.$$

Since the coefficients of $Q(x)$ have only one change in sign and since $Q(0)$ is negative, $Z(s)$ is truly a Subclass 4 function by Theorem 4. Therefore all the bounds given in Part II hold on the transient responses corresponding to this $Z(s)$.

In particular, the unit step response is

$$A(t) = .0295 - .0481e^{-t} + .0281e^{-2t} + .0247e^{-t} \cos(t + 1.89) \\ + .0064e^{-3t} \cos(3t + 4.43)$$

and this is plotted in Fig. 6. The bound, $t^4/1.413$, on the magnitude of $A(t)/r$, which is a consequence of Theorem 9, is shown by the dotted curves. Moreover pairs of values for the variables τ_s and δ are seen to be well above the curve for $m - n = 4$ in Fig. 4. For instance, for t greater than 2.60, $A(t)/r$ is bounded by $(1 \pm .129)$. Moreover the normalization factor T_N has a value of .429 for this example. The point determined by $\tau_s/T_N = 6.06$ and $\delta = .129$ is above the appropriate curve in Fig. 4.

3. A network [18], which has satisfactory responses for appropriate choices of the network parameters is shown in Fig. 7. Both the series-shunt peaking network and the

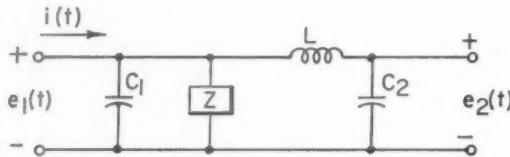


FIG. 7. A network whose transfer function of output voltage $E_2(s)$ divided by input current $I(s)$ is always a Subclass 3 function for all positive values of C_1 , C_2 , and L and all driving point impedances Z having a finite, non-zero value at DC .

Dietzold network have this form and their transient responses of the output voltage $e_2(t)$ as a function of the input current $i(t)$ are plotted in [18].

It is easily seen that if the driving point impedance $Z(s)$ has a finite non-zero value at DC , then the transfer function $Z_T(s)$ of output voltage $E_2(s)$ divided by input current $I(s)$ will always be a Subclass 3 function for all positive values of C_1 , C_2 , and L and for all permissible driving point impedances $Z(s)$. For, the ratio of the output voltage to input voltage $E_1(s)$ is

$$\frac{E_2}{E_1} = \frac{1}{LC_2 s^2 + 1}.$$

But since $E_1(s) = I(s)Z_D(s)$ where $Z_D(s)$ is the input impedance of this network,

$$Z_T(s) = \frac{E_2}{E_1} \cdot \frac{E_1}{I} = \frac{1}{LCs^2 + 1} Z_D(s).$$

Thus,

$$R_T(\omega) = \operatorname{Re}[Z_T(j\omega)] = \frac{R_D(\omega)}{1 - LC\omega^2}.$$

Since $R_D(\omega)$ is the real part of the passive driving point impedance $Z_D(j\omega)$, it is non-

negative for all ω . And so, $R_T(\omega)$ will have exactly two changes of sign. This means that $Z_T(s)$ is a Subclass 3 function provided that $R_D(0)$ is non-zero and finite.

An inspection of the unit step responses corresponding to $Z_T(s)$ for the special cases of the series-shunt peaking network and the Dietzold network [18] shows that these transient responses satisfy all the restrictions developed in Part II of this paper. It should be noted that these networks are suitable for video amplifiers where a small rise time from 10 to 90 per cent coupled with small overshoot is of importance. If a small initial time delay is also desired, then these networks would not be particularly useful since they have Subclass 3 transfer impedances and their step responses must have appreciable initial time delays.

APPENDIX II

List of symbols. Those symbols which have physical significance are defined as follows:

$A(t)$,	the response to a unit step function applied at time, $t = 0$;
a_k ,	the coefficients in the numerator of a system function;
b_k ,	the coefficients in the denominator of a system function;
$D(s)$,	the denominator of a system function;
$f_t(x)$,	the Sturm functions defined in Theorem 6;
$I(\omega)$,	the imaginary part of a system function for real frequencies;
$I_a(\omega)$,	a function defined by Eq. (4);
K ,	the constant multiplier of a system function;
m ,	the degree of the denominator of a system function;
$N(s)$,	the numerator of a system function;
n ,	the degree of the numerator of a system function;
$P(\omega^2)$,	the even part of $N(j\omega) D(-j\omega)$;
$Q(\omega^2)$,	the even part of $N(j\omega) D(-j\omega)/\omega$;
$R(\omega)$,	the real part of a system function for real frequencies;
$R_a(\omega)$,	a function defined by Eq. (3);
r ,	the resistance of a system function under DC conditions;
s ,	the complex frequency variable;
T_N ,	a time normalization factor defined by expression (37);
T_γ ,	the rise time from zero to one of the step response;
t ,	the time variable;
$W(t)$,	the response to a unit impulse function applied at time, $t = 0$;
x ,	the square of angular velocity = ω^2 ;
$Z(s)$,	a system function;
$Z_a(s)$,	a function defined by Eq. (2);
γ ,	the least upper bound on all the fractional overshoots and undershoots of the unit step response;
δ ,	the least upper bound on $ A(t) - r /r$ for $t \geq \tau_i$;
θ_i ,	the phase angle for a pole factor defined in Fig. 1;
μ_i ,	a zero of a system function;
ρ_i ,	a pole of a system function;
σ ,	the real part of the complex frequency variable;
τ ,	any time beyond which the unit step response remains within the bounds $(1 \pm \gamma)r$;

- τ_δ , the least time beyond which the unit step response remains within the bounds $(1 \pm \delta)r$ where $0 \leq \delta \leq 1$;
 $\varphi(\omega)$, the phase angle of a system function defined by Eq. (14);
 Ψ_i , the phase angle for a zero factor defined in Fig. 1;
 ω , the imaginary part of the complex frequency variable, s .

Acknowledgment. This investigation was supported by a grant from the research reserve of the College of Engineering, New York University.

REFERENCES

1. A. H. Zemanian, *Bounds existing on the time and frequency responses of various types of networks*, Proc. IRE 42, 835-839 (May 1954)
2. A. H. Zemanian, *Further bounds existing on the transient responses of various types of networks*, Proc. IRE 43, 322-326 (March 1955)
3. A. H. Zemanian, *Restrictions on the shape factors of the step response of positive real system functions*, Proc. IRE, 44, 1160-1165 (Sept. 1956)
4. I. A. Ovseyevich, *Certain bounds on the time functions of a linear system given by its frequency characteristic*, Izvestia Akad. Nauk, Otdel. Tekh. Nauk, S.S.R., pp. 59-68 (Feb. 1956)
5. O.P.D. Cuttridge, *Transient response of two-terminal networks*, Inst. Elec. Eng., Monograph No. 212R (Dec. 1956)
6. E. W. Tschudi, *Admittance and transfer function for an n-mesh R C filter network*, Proc. I.R.E. 38, 309-310 (March 1950)
7. R. R. Kenyon, *Response characteristics of resistance-reactance ladder networks*, Proc. I.R.E. 39, 557-559 (May 1951)
8. L. Storch, *The multisection RC filter network problem*, Proc. I.R.E. 39, 1456-1458 (Nov. 1951)
9. B. K. Bhattacharya, *Admittance and transfer function of a multimesh resistance-capacitance filter network*, Indian J. Phys. 26, 563-574 (Nov. 1952)
10. E. Green, *Amplitude-frequency characteristics of ladder networks*, Marconi's Wireless Telegraph Co., Chelmsford, Essex, 1954, pp. 6-11
11. E. A. Guillemin, *The mathematics of circuit analysis*, John Wiley and Sons, New York, 1949, p. 400
12. *Ibid.*, p. 395
13. L. Weisner, *Introduction to the theory of equations*, The Macmillan Co., 1938, Chap. 5
14. A. Hurwitz, *Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt*, Math. Ann. 46, 273-284 (1895)
15. A. H. Zemanian, *Some inequalities for Fourier transforms*, Proc. Am. Math. Soc. 8, No. 3, (June 1957)
16. A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher transcendental functions*, vol. I, McGraw-Hill Book Co., 1953, Eq. 2.8 (46) and 1.2 (8)
17. H. T. Davis, *Tables of the higher mathematical functions*, vol. I, Principia Press, Bloomington, Ind., 1933
18. R. C. Palmer and L. Mautner, *A new figure of merit for the transient response of video amplifiers*, Proc. IRE 37, 1073-1077 (Sept. 1949)

**SOME NEW TECHNIQUES IN THE
DYNAMIC PROGRAMMING SOLUTION OF VARIATIONAL PROBLEMS***

BY

RICHARD BELLMAN

The Rand Corporation, Santa Monica, Cal.

Summary. In previous papers, it has been shown that the functional equation technique of dynamic programming may be applied to yield the numerical solution of a wide class of variational problems of the type occurring in mathematical physics, engineering, and economics.

It was seen that the numerical solution of a problem involving N state variables depended upon the computation of sequences of functions of N variables. This fact made the method routine only for the case where $N = 1$ or 2, with grave difficulties arising in the general case.

In this paper, it is indicated how this difficulty can be overcome for a large class of problems in which the underlying equations and criterion function are linear, although the restraints on the forcing functions may be non-linear, corresponding say to energy considerations.

The same methods are applicable to other classes of linear equations, and, in particular, to differential-difference equations, arising from time-lag problems and to various classes of partial differential equations. These problems could not previously be treated by dynamic programming techniques in any usable fashion.

Finally, it is briefly indicated how the method of successive approximations may be combined with the foregoing techniques to reduce general variational problems with non-linear equations and criterion function to sequences of problems that can be solved numerically by means of sequences of functions of one variable. There are a number of interesting and difficult convergence questions associated with this program; these, however, are not discussed here.

1. Introduction. A variational problem that is encountered in many areas of pure and applied mathematics is that of determining the minimum or maximum of a functional of the form

$$J(y) = \int_0^T F(x_1, x_2, \dots, x_N; y_1, y_2, \dots, y_M) dt \quad (1.1)$$

over all functions y_1, y_2, \dots, y_M , connected to the x_i by means of the relations

$$\frac{dx_i}{dt} = G_i(x_1, x_2, \dots, x_N; y_1, y_2, \dots, y_M), \quad x_i(0) = c_i, \quad i = 1, 2, \dots, N, \quad (1.2)$$

and satisfying constraints of the form

$$(a) \quad x_i(T) = b_i, \quad i = 1, 2, \dots, k, \\ (b) \quad R_j(x; y) \leq 0, \quad j = 1, 2, \dots, m. \quad (1.3)$$

For a variety of reasons, which preliminary mathematical study discloses quite readily, this problem, to the degree of generality stated above, presents formidable analytic

*Received Sept. 23, 1957.

difficulties. Not only do these difficulties arise in connection with the explicit solution of the problem, see [9] and [10], but also in connection with the apparently more modest demand for a numerical solution.

We have discussed elsewhere, [1, 2, 3, 4], various applications of the functional equation technique of dynamic programming to the numerical solution of classes of problems of the foregoing type. In this paper, we wish to indicate some recent developments which greatly enlarge the scope of the methods utilized in the references cited above. These new developments, combined with the classical tool of successive approximations, enable us to attack systematically large classes of problems formerly far beyond our powers.

We shall begin our discussion with a terminal control problem for a linear system with constant coefficients. The problem we consider is that of maximizing a functional of the form

$$J(y) = H[x_1(T), x_2(T), \dots, x_k(T)] \quad (1.4)$$

over all functions y_i related to the x_i by means of the linear differential equations

$$\frac{dx_i}{dt} = \sum_{j=1}^N a_{ij}x_j + \sum_{j=1}^N b_{ij}y_j, \quad x_i(0) = c_i, \quad i = 1, 2, \dots, N, \quad (1.5)$$

and satisfying constraints of the following type:

- (a) $m_i \leq y_i(t) \leq m'_i, \quad 0 \leq t \leq T, \quad i = 1, 2, \dots, N,$
 - (b) $\int_0^T K_j(y_1, y_2, \dots, y_N) dt \leq c_j, \quad j = 1, 2, \dots, l.$
- (1.6)

Whereas the techniques previously given enable us to convert this problem into one involving the sequential computation of functions of N variables, the linearity of the defining equation in (1.5) permits, as we shall see, a transformation of the problem into one involving sequences of functions of k variables, where k is as in (1.4).

Since current digital computers do not allow the storage of functions of more than two variables in any feasible way, this is a very important reduction if k is equal to one or two. As it turns out, a number of significant problems arising in the fields of economic and engineering control may be formulated in these terms.

If the function $H(x_1, x_2, \dots, x_k)$ is linear,

$$H(x_1, x_2, \dots, x_k) = \sum_{i=1}^k \alpha_i x_i, \quad (1.7)$$

the problem may be still further reduced to a computation involving sequences of functions of one variable, regardless of the values of k and N . The same result holds for the case where it is desired to maximize a linear functional of the form

$$J = \int_0^T \left[\sum_{i=1}^N \rho_i x_i(t) \right] dt. \quad (1.8)$$

In exactly the same way, we can treat discrete variational problems where difference equations replace differential equations. As a matter of fact, these techniques were developed in connection with a discrete problem, the "caterer" problem, [5].

As we shall see, the same methods are applicable to the treatment of variational

questions involving time lags and retarded control, see [7]. In simple cases involving only one time delay, the equation corresponding to (1.5) is

$$\frac{dx_i}{dt} = \sum_{j=1}^N a_{ij}x_j(t) + \sum_{j=1}^N \alpha_{ij}x_j(t-\delta) + \sum_{j=1}^N b_{ij}y_j(t), \quad (1.9)$$

$$x_i(t) = c_i(t), \quad 0 \leq t \leq \delta, \quad i = 1, 2, \dots, N.$$

More complicated types of hereditary processes and various classes of partial differential equations may also be treated by means of the techniques we present below. Formerly, problems of this nature could not be treated computationally by dynamic programming techniques because of their dependence upon functionals rather than functions.

As noted above, the results discussed in the foregoing paragraphs depend in an essential manner upon the linearity of the equations defining the process. To extend these techniques to cover more general processes, we turn to that general factotum of analysis, the method of successive approximations. Without entering into any of the questions of rigor, we indicate briefly how a variety of apparently multi-dimensional problems can be reduced to sequences of problems involving functions of one variable.

Although a large number of interesting and important questions concerning convergence, rapidity of convergence, stability, and so on, arise from these investigations, we shall postpone any investigation of these matters until a later date.

Similar techniques can be used to treat processes involving random effects. These will also be discussed subsequently.

2. Dynamic programming. In order to appreciate the improvement in technique afforded by the methods we present here, let us sketch briefly the direct approach of dynamic programming to variational problems of the aforementioned variety.

To treat the general question posed in (1.1)-(1.3), we define a function of N variables, $f(c_1, c_2, \dots, c_N; T)$, by means of the relation

$$f(c_1, c_2, \dots, c_N; T) = \max_y J(y), \quad (2.1)$$

where the maximum is taken over all functions $y_i(t)$ satisfying (1.3). Under appropriate assumptions concerning the continuous dependence of the maximizing functions upon the initial values c_i and the upper limit T , we obtain for the function f a non-linear partial differential equation

$$\frac{\partial f}{\partial T} = \max_{v_i} \left[F(c, v) + \sum_{i=1}^N G_i(c, v) \frac{\partial f}{\partial c_i} \right]$$

$$f(c_1, c_2, \dots, c_N; 0) \equiv 0.$$

The maximization in the above expression is over quantities v_i satisfying the constraints

$$(a) \quad m_i \leq v_i \leq m'_i, \quad i = 1, 2, \dots, N, \quad (2.3)$$

$$(b) \quad R_i(c, v) \leq 0.$$

If F and G satisfy suitable differentiability conditions, this non-linear partial differential equation leads, in the absence of constraints, to the usual Euler equations, via the method of characteristics, see [1, 2].

In general, in questions of greatest interest in applications, this reduction from partial differential equations to ordinary differential equations does not occur due to the presence of various constraints of a physical nature.

3. Computational aspects. Since the analytic solution of problems of the type discussed in the foregoing sections is only rarely achieved, we turn to the study of computational techniques. Our aim initially is to provide a numerical algorithm, based upon the use of a digital computer, which resolves questions of this nature in a routine fashion without the use of an extensive analysis strongly dependent upon the particular analytic structure of the functions occurring. In practice, simplifications can always be made, taking advantage of the structure of the particular problem under consideration. As we shall see below, it is only by proceeding in this manner that we can make significant advances. General methods, however, are always useful, if only as a court of last resort.

To determine $f(c_1, c_2, \dots, c_N, T)$, we can either use (2.2) and any of a number of standard techniques for the numerical solution of partial differential equations of this type, or, as has turned out to be preferable, we can go over to a discrete version of the original continuous process. Analytically, this means that the original differential equations are replaced by difference equations. Thus, (1.2) becomes

$$\begin{aligned} x_i(t + \delta) &= x_i(t) + \delta G_i[x_1(t), x_2(t), \dots, x_N(t); y_1(t), y_2(t), \dots, y_M(t)], \\ x_i(0) &= c_i, \quad i = 1, 2, \dots, N, \end{aligned} \quad (3.1)$$

with t assuming only the values $0, \delta, 2\delta, \dots$.

The non-linear partial differential equation is then replaced by the non-linear recurrence relation

$$\begin{aligned} f(c, T) &= \max_v [\delta F(c, v) + f[c_1 + \delta G_1(c, v), \dots, c_N + \delta G_N(c, v); T - \delta]], \\ f(c, 0) &\equiv 0. \end{aligned} \quad (3.2)$$

In order to carry out the indicated process, we must be able to tabulate functions of N variables. This means that at the present time the approach outlined above is only feasible if $N = 1$ or 2 . For $N > 2$, the memory requirements become prohibitive.

It is necessary then to develop some new techniques if we wish to utilize dynamic programming to solve large scale problems of the type arising in the engineering and economic spheres.

4. Preliminaries on linear systems. In this section we shall mention some well-known results concerning the solution of vector-matrix systems of linear differential equations. These will be utilized in what follows. Proofs of the results cited here may be found in [6].

The linear system of (1.5) may be written, using an obvious vector-matrix notation, in the form

$$\frac{dx}{dt} = Ax + By, \quad x(0) = c. \quad (4.1)$$

Consider first the case in which A is constant. The solution of (1) may then be written in the form

$$x = \exp(At)c + \int_0^t \exp[A(t-s)]By(s) ds. \quad (4.2)$$

If $A = A(t)$, a matrix dependent upon t , then x may be written in the form

$$x = X(t)c + \int_0^t X(t)X^{-1}(s)By(s) ds, \quad (4.3)$$

where $X(t)$ is the matrix solution of

$$\frac{dX}{dt} = A(t)X, \quad X(0) = I. \quad (4.4)$$

We shall utilize these representations in a crucial manner below.

5. Terminal control—General non-linear criterion. Let us now turn to the problem of maximizing a given function $H[x_1(T), x_2(T), \dots, x_k(T)]$ of the terminal state of the system over all control functions $y_i(t)$ which are related to the x_i by means of the linear equation in (4.1), and which are subject to the constraints

- (a) $m_i \leq y_i(t) \leq m'_i, \quad 0 \leq t \leq T, \quad i = 1, 2, \dots, N,$
- (b) $\int_0^T G(y_1, y_2, \dots, y_N) dt \leq k.$

We wish to show that the numerical solution of a problem of this type can be made to depend upon a sequence of functions of k variables, rather than upon sequences of functions of N variables. We shall consider first the case where $A = (a_{ij})$ is a constant matrix.

We begin with the linear representation of (4.2), which yields a set of equations

$$x_i(t) = z_i(t) + \int_0^t \left[\sum_{j=1}^N x_{ij}(t-s)y_j(s) \right] ds, \quad i = 1, 2, \dots, N, \quad (5.2)$$

where $z_i(t)$ is the i th component of $\exp(At)c$, and $X(t) = [x_{ij}(t)]$.

The problem we wish to consider may then be cast in the form of maximizing a functional of the type

$$H\left(u_1 + \int_0^T \left[\sum_{i=1}^N x_{1i}(T-s)y_i(s) \right] ds, \dots, u_k + \int_0^T \left[\sum_{i=1}^N x_{ki}(T-s)y_i(s) \right] ds \right), \quad (5.3)$$

where u_1, u_2, \dots, u_k are given quantities, over all functions y_1, y_2, \dots, y_N satisfying the constraints (5.1a) and (5.1b).

Let us then consider the sequence of functions $f(u_1, u_2, \dots, u_k; T)$, implicitly dependent upon λ , defined as follows.

$$f(u_1, u_2, \dots, u_k; T) = \max_u \left[H\left(u_1 + \int_0^T [\dots] ds, \dots, u_k + \int_0^T [\dots] ds \right) - \lambda \int_0^T G(y_1, y_2, \dots, y_N) ds \right], \quad (5.4)$$

where the functions $y_i(t)$ are now constrained only by (5.1a).

A motivation and discussion of this use of the Lagrange multiplier may be found in [8], and a numerical example in [11]. The value of the method lies in the fact that it enables us to reduce multi-dimensional problems, where the dimensionality is reckoned in the dynamic programming sense, to sequences of lower dimensional problems.

To obtain a functional equation for $f(u_1, u_2, \dots, u_k; T)$, we proceed as follows. Suppose that the values of $y_1(t), y_2(t), \dots, y_N(t)$ have been determined over $[0, \delta]$.

Then we may write

$$\begin{aligned}
 & H\left(u_1 + \int_0^T [\dots] ds, \dots, u_k + \int_0^T [\dots] ds\right) - \lambda \int_0^T G(y_1, y_2, \dots, y_N) ds \\
 & = H\left(u_1 + \int_0^\delta [\dots] ds + \int_\delta^T [\dots] ds, \dots, u_k + \int_0^\delta [\dots] ds + \int_\delta^T [\dots] ds\right) \\
 & \quad - \lambda \int_0^\delta G(y_1, y_2, \dots, y_N) ds - \lambda \int_\delta^T G(y_1, y_2, \dots, y_N) ds \\
 & = H\left(u_1 + \int_0^\delta [\dots] ds + \int_0^{T-\delta} \left[\sum_{i=1}^N x_{ii}(T - \delta - s)y_i(s + \delta) \right] ds, \dots\right) \\
 & \quad - \lambda \int_0^\delta G(y_1, y_2, \dots, y_N) ds - \lambda \int_0^{T-\delta} G[y_1(s + \delta), \dots, y_N(s + \delta)] ds.
 \end{aligned} \tag{5.5}$$

The principle of optimality, see [1], then yields the functional equation

$$\begin{aligned}
 f(u_1, u_2, \dots, u_k; T) &= \max_{v[0, \delta]} \left[-\lambda \int_0^\delta G(y_1, y_2, \dots, y_N) ds \right. \\
 &\quad \left. + f\left(u_1 + \int_0^\delta \left[\sum_{i=1}^N x_{ii}(T - s)y_i(s) \right] ds, \dots\right) \right].
 \end{aligned} \tag{5.6}$$

The maximum is now taken over all functions $y_i(s)$ defined over $0 \leq s \leq \delta$, and satisfying the constraints $m'_i \leq y_i(s) \leq m_i$ in $[0, \delta]$.

For computational purposes, we may use the approximate relation

$$\begin{aligned}
 f(u_1, u_2, \dots, u_k; T) &= \max_v \left[-\lambda \delta G(v_1, v_2, \dots, v_N) \right. \\
 &\quad \left. + f\left[u_1 + \delta \sum_{i=1}^N x_{ii}(T)v_i, \dots\right]\right]
 \end{aligned} \tag{5.7}$$

or we may start with a discrete version of the original process.

We have thus reduced the numerical solution of the variational problem to the determination of a sequence of functions of k variables. If $k = 1$ or 2, we have a feasible method of solution.

6. Terminal control—Quadratic criterion. Let us note that the problem of minimizing the functional

$$Q[x_1(T), x_2(T), \dots, x_N(T)] \tag{6.1}$$

over all functions y satisfying the relations

$$\frac{dx}{dt} = A(t)x + B(t)y, \quad x(0) = c, \tag{6.2}$$

can be reduced to the solution of systems of linear equations, if Q is a quadratic form in the $x_i(T)$; see [9].

The same holds if we add to Q a quadratic function of the form

$$\int_0^T P(x_1, x_2, \dots, x_N; y_1, y_2, \dots, y_N) dt, \tag{6.3}$$

where P is quadratic in its arguments.

Problems of this type may also be very simply treated by means of the formalism of dynamic programming, a matter we will discuss elsewhere.

7. Terminal control—Variable coefficients. It is important, in connection with our subsequent discussion of the use of successive approximations, to consider the same problem for the case where A is a variable matrix. Let us consider then the situation where the equation governing the process has the form

$$\frac{dx}{dt} = A(t)x + B(t)y + \phi(t), \quad x(0) = c. \quad (7.1)$$

As we know, the solution of this equation is given by the expression

$$x = x(t)c + \int_0^t x(s)x^{-1}(s)B(s)y(s) ds + \int_0^t x(s)x^{-1}(s)\phi(s) ds. \quad (7.2)$$

Hence the components of $x(T)$ have the form

$$x_i(T) = u_i + \int_0^T \left[\sum_{i=1}^N x_{ii}(T, s)y_i(s) \right] ds, \quad (7.3)$$

where the u_i are independent of y .

In order to take account of the non-stationarity of the process, we count time backwards. In place of noting the time at which the process ends, we single out the time at which it begins. Fixing T , we consider the function $f(u_1, u_2, \dots, u_k; r)$ defined by the relation

$$f(u_1, u_2, \dots, u_k; r) = \max_{v(r, T)} \left[H \left[u_1 + \int_r^T \left[\sum_{i=1}^N w_{ii}(T, s)y_i(s) ds, \dots \right] - \lambda \int_r^T G(y_1, y_2, \dots, y_N) ds \right] \right] \quad (7.4)$$

Arguing as in the preceding section, we see that f satisfies the relation

$$f(u_1, u_2, \dots, u_k; r) = \max_{v(r, r+\delta)} \left[-\lambda \int_r^{r+\delta} G(y_1, y_2, \dots, y_N) ds + f \left(u_1 + \int_r^{r+\delta} [\dots] ds, \dots, u_k + \int_r^{r+\delta} [\dots] ds \right) \right] \quad (7.5)$$

For computational purposes, this reduces to

$$f(u_1, u_2, \dots, u_k; r) = \max_v \left[-\lambda \delta G(v_1, v_2, \dots, v_N) + f \left[u_1 + \delta \sum_{i=1}^N w_{ii}(T, r)v_i, \dots \right] \right], \quad (7.6)$$

with $f(u_1, u_2, \dots, u_k; T) = 0$. Here $r = n\delta, \dots, T = M\delta$.

8. Terminal control—Linear criterion. Let us now consider the case where H is a linear function, which is to say, we consider the problem of maximizing the inner product $[s(T), a]$ where a is a given vector. To simplify the notation, let us consider only the case where A is a constant matrix.

Using the representation for $x(t)$ given in (4.2), we see that

$$[x(T), a] = (\exp(AT)c, a) + \left[\int_0^T \exp[A(T-s)]y(s) ds, a \right]. \quad (8.1)$$

Neglecting the term $[\exp(AT)c, a]$, which is independent of y , we have the problem of maximizing

$$J(y) = \left[\int_0^T \exp[A(T-s)]y(s) ds, a \right] \quad (8.2)$$

over all y_i satisfying the constraints

$$\begin{aligned} (a) \quad m_i &\leq y_i \leq m'_i, \quad 0 \leq t \leq T, \quad i = 1, 2, \dots, N, \\ (b) \quad \int_0^T G(y_1, y_2, \dots, y_N) ds &\leq k. \end{aligned} \quad (8.3)$$

Introduce the function

$$f(k, T) = \max_y J(y), \quad (8.4)$$

where the maximum is over all functions y_i satisfying (8.3a) and (8.3b), and k is as in (8.3b).

It is easy reasoning as above to see that $f(k, T)$ satisfies the equation

$$\begin{aligned} f(k, T) &= \max_{\nu(0, \delta)} \left[\left(\int_0^\delta \exp[A(T-s)]y ds, a \right) \right. \\ &\quad \left. + f \left[k - \int_0^\delta G(y_1, y_2, \dots, y_N) ds, T - \delta \right] \right], \quad f(k, 0) \equiv 0. \end{aligned} \quad (8.5)$$

For computational purposes, we can use the relation

$$f(k, T) = \max_v [\delta(\exp(AT)v, a) + f(k - \delta G(v_1, v_2, \dots, v_N), T - \delta)]. \quad (8.6)$$

We see then that in the case where the underlying equation is linear, and there is only one constraint of the form in (8.3b), we can compute the solution using sequences of functions of one variable.

If there are two constraints of the type

$$\int_0^T G_i(y_1, y_2, \dots, y_N) ds \leq k_i, \quad i = 1, 2. \quad (8.7)$$

we introduce a Lagrange multiplier and consider the problem of maximizing

$$[x(T), a] - \lambda \int_0^T G_1(y_1, y_2, \dots, y_N) dt. \quad (8.8)$$

For each value of λ we have a one-dimensional problem. As the parameter λ is varied, we range over a set of values of the constraint $\int_0^T G_1(y_1, y_2, \dots, y_N) ds$.

In some cases, if the constraint is that given in (8.3b), we can solve the problem analytically, see [9, 10], and thus eliminate all computational aspects. It can easily

happen that a direct numerical solution for a range of values of k and c consumes less time and effort than a solution based upon an explicit analytic expression.

9. Time-lag processes. As the preceding sections show, the success of the method presented in the foregoing sections rested upon the superposition principle. Given an equation of the form

$$\begin{aligned} L(x) &= y, \\ x_{t=0} &= c, \end{aligned} \tag{9.1}$$

we were able to write the solution in the form

$$x = X(t)c + T(y), \tag{9.2}$$

thus avoiding any interaction between the initial state and the forcing function.

Also important was the fact that $T(y)$ had the form

$$T(y) = \int_0^t K(t, s)y(s) ds. \tag{9.3}$$

It follows from this analysis, that the methods of the preceding sections are applicable to situations in which the underlying equations are differential-difference equations of the form given in (1.9). For the case where the coefficients are constant, Laplace transform methods yield the requisite representation formulas, see [7]. For the case of variable coefficients, these representation formulas have been developed in a forthcoming paper by the author and K. Cooke.

10. Heat conduction processes. In the study of the control of thermal processes, and in connection with the recent field of nuclear reactor control, we encounter variational problems in which the underlying equation is

$$\begin{aligned} u_t - u_{xx} &= f(x, t), \quad 0 < x < 1, \quad t > 0, \\ u(x, 0) &= c(x), \quad 0 \leq x \leq 1, \\ u(0, t) &= u(1, t) = 0, \quad t > 0. \end{aligned} \tag{10.1}$$

Since the solution of this equation possesses the requisite properties described in the preceding section, it follows that a number of variational problems involving this equation can be treated by means of the foregoing techniques. A detailed discussion will appear subsequently.

11. Successive approximations—I. Let us now briefly, without entering into any rigorous discussion, which as may be imagined is non-trivial, indicate how the method of successive approximations may be combined with the foregoing techniques so as to reduce the computational solution of general variational problems involving non-linear differential equations to sequences of computational problems involving functions of fewer than N variables.

An important point to emphasize is that modern digital computers enable us to use as single steps in a computational process the solutions of problems once considered formidable in their own right.

Consider the problem of maximizing

$$J(y) = H[x_1(T), x_2(T), \dots, x_k(T)] \tag{11.1}$$

over all $y_i(t)$ subject to

- (a) $\frac{dx_i}{dt} = G_i(x, y), \quad x_i(0) = c_i, \quad i = 1, 2, \dots, N,$
 - (b) $m_i \leq y_i \leq m'_i, \quad i = 1, 2, \dots, N,$
 - (c) $\int_0^T G(y_1, y_2, \dots, y_N) dt \leq k.$
- (11.2)

Let $y^0(t) = [y_1^0(t), y_2^0(t), \dots, y_N^0(t)]$ be an initial guess in policy space, satisfying (11.2b) and (11.2c), and let $x^0(t) = [x_1^0(t), x_2^0(t), \dots, x_N^0(t)]$ be the set of x -values determined by (11.2a) when $y(t)$ is replaced by $y^0(t)$.

Consider the new system of differential equations

$$\begin{aligned} \frac{dx_i}{dt} &= G_i(x^0, y^0) + \sum_{j=1}^N (x_j - x_j^0) \frac{\partial G_i}{\partial x_j}(x^0, y^0) \\ &\quad + \sum_{j=1}^N (y_j - y_j^0) \frac{\partial G_i}{\partial y_j}(x^0, y^0) \quad i = 1, 2, \dots, N. \end{aligned} \quad (11.3)$$

The new variational problem is that of maximizing $J(y)$ over all y satisfying (11.2b), (11.2c), and related to x by means of (11.3).

Since the underlying system in (11.3) is linear, this variational problem may be treated in terms of functions of k variables. Let the maximizing $y(t)$ be called $y^1(t)$. Proceeding as above, we determine $x^1(t)$ using (11.2a), and consider the new approximating linear equation

$$\frac{dx_i}{dt} = G_i(x^1, y^1) + \sum_{j=1}^N (x_j - x_j^1) \frac{\partial G_i}{\partial x_j}(x^1, y^1) + \sum_{j=1}^N (y_j - y_j^1) \frac{\partial G_i}{\partial y_j}(x^1, y^1). \quad (11.4)$$

The new variational problem determines a vector y^2 which enables us to determine a new state vector x^2 by way of (11.2a). Continuing in this way, we obtain a sequence of vector functions $\{y^k\}$, and a sequence of state vectors $\{x^k\}$, which we hope converges to a solution to the original variational problem.

12. Successive approximations—II. If the criterion function $H[x_1(T), x_2(T), \dots, x_k(T)]$ is linear,

$$H(x_1, x_2, \dots, x_k) = (x, b), \quad (12.1)$$

then the approximation procedure outlined above yields a problem which at each step can be resolved computationally in terms of sequences of functions of one variable, and even in explicit analytic terms under favorable circumstances.

13. Terminal control vs. general control. We have placed considerable emphasis upon terminal control processes because of the fact that a simple transformation enables us to treat general control processes as terminal control processes. If

$$J(y) = \int_0^T F(x_1, x_2, \dots, x_N; y_1, y_2, \dots, y_N) dt, \quad (13.1)$$

the introduction of a new variable, $x_{N+1}(t)$, determined by the relation

$$\frac{dx_{N+1}}{dt} = F(x_1, x_2, \dots, x_N; y_1, y_2, \dots, y_N), \quad x_{N+1}(0) = 0, \quad (13.2)$$

yields a new problem in which we wish to maximize $x_{N+1}(T)$.

Similarly, if we have the criterion function $G[x_1(T), x_2(T), \dots, x_N(T)]$, a new variable, $x_{N+1}(t)$, determined by the relation

$$\frac{dx_{N+1}}{dt} = \sum_{i=1}^N \frac{\partial G}{\partial x_i} \frac{dx_i}{dt} = \sum_{i=1}^N \frac{\partial G}{\partial x_i} H_i(x, y), \quad (13.3)$$

once again reduces the variational problem to one of terminal control.

BIBLIOGRAPHY

1. R. Bellman, *Dynamic programming*, Princeton University Press, 1957
2. R. Bellman, *Dynamic programming and its application to variational problems in mathematical economics*, Proceedings Symposium on Calculus of Variations and its Applications, Am. Math. Soc., 1955
3. R. Bellman, *On the application of the theory of dynamic programming to the study of control processes*, Symposium on Non-linear circuit analysis, Polytechnic Institute of Brooklyn, vol. VI, 1956
4. R. Bellman, *Notes on the theory of control processes—I: On the minimum of maximum deviation*, Quart. Appl. Math. **XIV**, 419–423 (1957)
5. R. Bellman, *On a dynamic programming approach to the caterer problem—I*, Management Sci. **3**, 270–278 (1957)
6. R. Bellman, *Stability theory of differential equations*, McGraw-Hill, 1953
7. R. Bellman, *A survey of the mathematical theory of time-lag, retarded control, and hereditary processes*, The RAND Corp. Rept. R-256, 1954
8. R. Bellman, *Dynamic programming and Lagrange multipliers*, Proc. Natl. Acad. Sci. **42**, 767–769, (1956)
9. R. Bellman, I. Glicksberg and O. Gross, *On some variational problems occurring in the theory of dynamic programming*, Rendiconti del Circolo Matematico di Palermo, Serie II, Tomo **III** 1–35 (1954)
10. R. Bellman, W. H. Fleming, and D. V. Widder, *Variational problems with constraints*, Annali di Matematica, Serie IV, Tomo **XLI**, 301–323, 1956
11. S. Dreyfus, *Dynamic programming solution of allocation problems*, Techniques of Industrial Operations Research Seminar, Illinois Institute of Technology, June 1956
12. H. Osborn, *Euler equations and characteristics*, Chap. 7 of *Dynamic programming of continuous processes*, The RAND Corp., Rept. R-271, 1955

BOOK REVIEWS

(Continued from p. 258)

It is addressed to physicists, experimental as well as theoretical, who are familiar with the elements of quantum mechanics and who are approaching the subject of angular momentum for the first time.

Accordingly, it is not surprising that the author, in his presentation, avoids group-theoretical and abstract algebraic methods. Instead, he relies upon an inductive approach to his subject which leans rather heavily on the assumption that the reader, from previous experience, already has an intuitive idea as to what is meant by "angular momentum" and "spin." There is an inconsistency here in this assumption which may not be out of place in a spoken lecture but which some, like the reviewer, may find annoying to see in a more formal treatment of the subject.

The book is divided into two parts of approximately equal length, entitled General Theory and Applications. Under General Theory, one finds a terse review of some basic principles of quantum mechanics (which the reviewer feels could well have been omitted), a discussion of the angular-momentum operators (these are defined in terms of the transformation properties of wave functions under rotations), together with chapters dealing with the addition of two angular momenta and the Clebsch-Gordan coefficients, the transformation properties of the angular-momentum wave functions under rotations, irreducible tensors and the Wigner-Eckart theorem, and, finally, the addition of three angular momenta and the Racah coefficients. Applications include chapters on the expansion of the electromagnetic field into multipole fields, the multipole moments of static charge distributions, spin one-half particles, oriented nuclei and angular correlations, angular distributions in nuclear reactions, and wave functions for systems of identical particles.

The discussion of the general theory is sufficiently detailed and up to date so as to give the reader a useful first approach to the theory of angular momentum. The applications are sufficiently varied so as to illustrate the considerable scope of the subject; it is inevitable, however, that the author cannot go into as much detail here as in the discussion of the general theory. It is for this reason that the second half of the book does not hang together nearly so well as the first part; it can, however, serve as a very useful point of departure to the periodical literature.

Since it seems fairly certain that this book will be widely used as a first introduction to the subject, one would have hoped that certain inconsistencies and ambiguities, which may trouble the beginner, had been avoided. For example, on p. 16, there is the remark that, in quantum mechanics, "no more than one component operator (of the angular momentum) can be a constant of the motion." On p. 81, we have the statement that "a scalar, as the term is used here, may be a tensor of any rank." Finally, the author consistently uses the term "projection quantum number" in place of the old-fashioned "magnetic quantum number" except for the several instances where he forgets himself and reverts to the more conventional nomenclature.

DAVID FELDMAN

Applied group-theoretic and matrix methods. By Bryan Higman. Oxford University Press, New York (American Branch), 1955. xii + 454 pp. \$9.60.

This book is an outgrowth of lectures given at the University College of the Gold Coast, the object being to present the theory of matrices and group representations in a form palatable to chemists and physicists. The style of the book, which is quite unique, reveals its origin in a rather informal lecture series and the author probably succeeds in conveying the intuitive picture that goes along with the formal theory. The author, however, shows excessive disregard for mathematical rigor to the point of making rather vague statements of theorems. Most of the book is devoted to the applications in almost all the obvious branches of physics and chemistry. At places these are discussed in monotonous detail and one is left with the impression that the author, in amplifying his lecture notes, did not know where to stop.

G. F. NEWELL

(Continued on p. 318)

—NOTES—

EXTENSION OF MICHELL'S THEOREM TO PROBLEMS OF PLASTICITY AND CREEP*

BY BERNARD BUDIANSKY (*Harvard University*)

A well known theorem of linear, isotropic elasticity due to J. H. Michell [1] gives the conditions under which the generalized plane stress distribution in a multiply-connected sheet subjected to prescribed boundary stresses is independent of Poisson's ratio. Such independence exists if, and only if, the resultant force (not necessarily the couple) on each boundary vanishes. It is shown in this paper that the same independence of the elastic value of Poisson's ratio holds under the same conditions even when plastic flow and creep occur¹.

Let $\sigma_{\alpha\beta}(x, y; t)$, $\epsilon_{\alpha\beta}(x, y; t)$, and $u_\alpha(x, y; t)$ be the time dependent stress, strain, and single-valued displacement distributions² that constitute a solution to the following time-dependent boundary value problem for the region R bounded externally by the curve C_0 and internally by the curves C_i ($i = 1, 2, \dots, N$). In R :

$$\sigma_{\alpha\beta,\alpha} = 0, \quad (1)$$

$$\epsilon_{\alpha\beta} = \frac{1}{E} [(1 + \nu)\sigma_{\alpha\beta} - \nu\sigma_{\gamma\gamma}\delta_{\alpha\beta}] + f_{\alpha\beta}, \quad (2)$$

$$\epsilon_{\alpha\beta} = \frac{1}{2}(u_{\alpha,\beta} + u_{\beta,\alpha}), \quad (3)$$

On C_i :

$$\sigma_{\alpha\beta}n_\alpha^{(i)} = T_\beta^{(i)} \quad (i = 0, 1, 2, \dots, N). \quad (4)$$

In this formulation E is Young's modulus, ν is Poisson's ratio, $n_\alpha^{(i)}$ is the unit outward normal to the boundary C_i , and the $T_\beta^{(i)}$ are prescribed, time-dependent distributions of boundary traction³. The functions $f_{\alpha\beta}$ ($= f_{\beta\alpha}$) represent the plasticity and creep contributions to the strain, and are permitted to depend on time and the detailed history of stress. The stress-strain relation (2) therefore represents the most general (isothermal) plasticity and creep law for elastically isotropic materials.

Next, consider a different material obeying the same stress-strain law (2) with the exception that Poisson's ratio is replaced by $\nu^* \neq \nu$; the value of E , and the functional dependence of $f_{\alpha\beta}$ on time and stress history are assumed to remain unchanged. To determine whether or not $\sigma_{\alpha\beta}$ remains a solution of the same boundary value problem for the new material, it is only necessary to see whether the strains

$$\epsilon_{\alpha\beta}^* = \frac{1}{E} [(1 + \nu^*)\sigma_{\alpha\beta} - \nu^*\sigma_{\gamma\gamma}\delta_{\alpha\beta}] + f_{\alpha\beta} \quad (5)$$

are derivable by means of the strain-displacement equation (3) from some single-valued displacement u_α^* .

*Received August 12, 1957.

¹In the case of elasticity, independence of Poisson's ratio is equivalent (by dimensional analysis) to independence of all elastic constants; this is no longer true in the present case of plasticity and creep.

²The usual summation convention and subscript notation for tensors and vectors are used here, with Greek subscripts taking on the values 1, 2.

³Whether Michell's conditions are fulfilled or not, the $T_\beta^{(i)}$ must, of course, satisfy conditions of equilibrium of the body as a whole.

Now let

$$\epsilon'_{\alpha\beta} = \epsilon_{\alpha\beta} - \epsilon^*_{\alpha\beta}. \quad (6)$$

Then, from (2) and (5),

$$\epsilon'_{\alpha\beta} = \frac{\nu - \nu^*}{E} (\sigma_{\alpha\beta} - \sigma_{\gamma\gamma} \delta_{\alpha\beta}). \quad (7)$$

Next consider the integral

$$J = \frac{E}{\nu - \nu^*} \iint_R \epsilon'_{\alpha\beta} \tau_{\alpha\beta} dA, \quad (8)$$

where $\tau_{\alpha\beta}$ is any distribution of stress satisfying equilibrium:

$$\tau_{\alpha\beta,\alpha} = 0 \quad (9)$$

and producing zero boundary tractions:

$$\tau_{\alpha\beta} n^{(i)}_\alpha = 0 \quad \text{on } C_i \quad (i = 0, 1, 2, \dots, N). \quad (10)$$

Now, if $\epsilon'_{\alpha\beta}$ is derivable from a single-valued displacement, it follows from the principle of virtual work that the integral J given by (8) vanishes. But a converse theorem is also valid; if $J = 0$ for all $\tau_{\alpha\beta}$ satisfying (9) and (10), then $\epsilon'_{\alpha\beta}$ is derivable from a single-valued displacement [2, 3]⁴. Substituting (7) into (8) gives

$$J = \iint_R (\sigma_{\alpha\beta} \tau_{\alpha\beta} - \sigma_{\alpha\alpha} \tau_{\beta\beta}) dA. \quad (11)$$

The stresses $\tau_{\alpha\beta}$ can be related to a stress function ϕ by

$$\tau_{\alpha\beta} = \phi_{,\gamma\gamma} \delta_{\alpha\beta} - \phi_{,\alpha\beta} \quad (12)$$

and then (11) becomes

$$J = - \iint_R \sigma_{\alpha\beta} \phi_{,\alpha\beta} dA. \quad (13)$$

As a consequence of the boundary conditions (10) on $\tau_{\alpha\beta}$, it follows that $\phi_{,\alpha}$ is single valued, and, furthermore, has a constant value, say $K_\alpha^{(i)}$, on each boundary (see, for example, [5], p. 191). It is therefore possible to transform (13) as follows:

$$\begin{aligned} J &= - \iint_R [(\sigma_{\alpha\beta} \phi_{,\alpha})_{,\beta} - (\sigma_{\alpha\beta,\beta}) \phi_{,\alpha}] dA, \\ &= - \sum_{i=0}^N K_\alpha^{(i)} \oint_{C_i} \sigma_{\alpha\beta} n_\beta^{(i)} dS, \\ &= - \sum_{i=0}^N K_\alpha^{(i)} P_\alpha^{(i)} dS, \quad \text{where } P_\alpha^{(i)} = \oint_{C_i} T_\alpha^{(i)} dS. \end{aligned}$$

⁴The proof in [2] is specifically limited to simply-connected regions; the proof in [3], and the supporting theorems contained in [4], are valid for multiply-connected regions.

Hence, if $P_a^{(i)} = 0$ on each boundary, J vanishes for all admissible choices of $\tau_{\alpha\beta}$. Hence, by the converse theorem mentioned above, $\epsilon'_{\alpha\beta}$ is derivable from a single valued displacement. The same is then necessarily true of $\epsilon_{\alpha\beta}^* = \epsilon_{\alpha\beta} + \epsilon'_{\alpha\beta}$, and hence $\sigma_{\alpha\beta}$ remains a solution for the stress when Poisson's ratio is changed. On the other hand, if $P_a^{(i)}$ does not vanish on some boundaries, a suitable choice of $\tau_{\alpha\beta}$ can always be made to render J non-zero. But this would necessarily imply that the strains $\epsilon'_{\alpha\beta}$ (and hence $\epsilon_{\alpha\beta}^*$) are not derivable from a single valued displacement, whence $\sigma_{\alpha\beta}$ would certainly not constitute a solution for the new material.

The present theorem can be useful in simplifying the initial formulation of some problems. For example, the choice $\nu = \frac{1}{2}$ in conjunction with a total stress-strain law of plasticity permits the use of a single formula for the sum of the elastic and plastic components of strain; in other problems, the choice $\nu = 0$ might be more appropriate. In addition, the present theorem may conceivably have significance in connection with photoplasticity.

Acknowledgment. The financial support of the Office of Naval Research under Contract Nonr 1866(02) is gratefully acknowledged.

REFERENCES

1. J. H. Michell, *On the direct determination of stress in an elastic solid, with application to the theory of plates*, Proc. London Math. Soc. **31**, 100-124 (1899)
2. W. S. Dorn and A. Schild, *A converse to the virtual work theorem for deformable solids*, Quart. Appl. Math. **14**, No. 2 (July 1956)
3. Bernard Budiansky and Carl E. Pearson, *On variational principles and Galerkin's procedure for nonlinear elasticity*, Quart. Appl. Math. **14**, No. 3 (October 1956)
4. Bernard Budiansky and Carl E. Pearson, *A note on the decomposition of stress and strain tensors*, Quart. Appl. Math. **14**, No. 3 (October 1956)
5. S. Timoshenko and J. N. Goodier, *Theory of elasticity*, 2nd ed., McGraw-Hill Book Co., Inc., 1951

ON ISOPERIMETRIC INEQUALITIES IN PLASTICITY*

By WALTER SCHUMANN (Brown University)

Abstract. The purpose of this paper is the proof of the inequality $P \geq 6\pi M_0$, where P is the total limit load, M_0 the yield moment of a thin, perfectly plastic, simply supported, uniformly loaded plate of arbitrary shape and connection.

Introduction. The theory of thin, rigid-perfectly plastic plates, given by Hopkins and Prager [1]** has been applied to circular plates with various load and edge conditions. However, if one tries to extend this theory to non-symmetrical cases, serious difficulties arise in seeking examples of exact solutions, although some cases have been solved (see for instance [2]). As a contribution to the estimation of the limit load in an arbitrary plate we shall use here the isoperimetric inequality, which relates a circular domain to an arbitrary domain in a convenient manner. One of the principal theorems of limit analysis [3] and the methods for isoperimetric problems given in Polya's and Szegő's book [4] will be used. Similar problems have been proposed and solved for other physical quantities, as for example the torsional rigidity, the principal frequency, etc.

*Received August 16, 1957. The results presented in this paper were obtained in the course of research conducted under Contract Nonr 562(10) by the Office of Naval Research and Brown University.

**Numbers in square brackets refer to the bibliography at the end of the paper.

2. A lower bound for the limit load of a simply supported plate of arbitrary shape under uniform load. Let us consider a simply supported plate of a rigid-perfectly plastic material obeying Tresca's yield condition (Fig. 1), and let p be the limit load

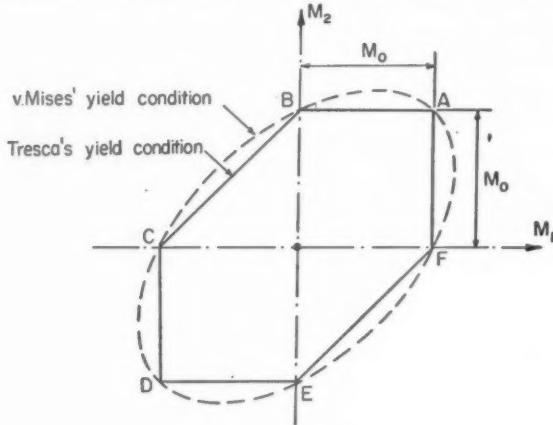


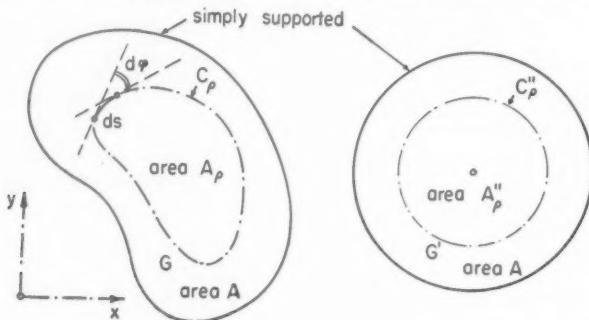
FIG. 1.

per unit area, which is assumed to be constant, A the area of the domain G (Fig. 2), $P = pA$ the total limit load and M_0 the yield moment. Then the following inequality is true:

$$P \geq 6\pi M_0 , \quad (1)$$

where the equality sign holds *only* in the case of the circular plate.

To prove this statement, we consider, in addition to the domain G , a circular domain G' of equal area A (Fig. 2), and we map the actual velocity field v of G into a new field v' over G' in a certain way, that will be defined later. All further quantities for the new

FIG. 2. Schwarz' symmetrization: $A_p = A_p''$.

field will be distinguished by primes from the corresponding quantities of the original field. Denoting by D the dissipation function per unit area, we have

$$\iint_{\sigma} p v \, dA = \iint_{\sigma} D \, dA. \quad (2)$$

Using the second theorem of limit analysis [3] and assuming that v' is kinematically admissible for G' , we have

$$\iint_{G'} p' v' dA' \leq \iint_{G'} D' dA'. \quad (3)$$

Suppose for the moment $v \geq 0$, and let D_{tot} be the total rate of dissipation, V the volume between the plane of the plate and the surface into which the plate deforms; then (2) and (3) give

$$P = \frac{D_{\text{tot}}}{V} A, \quad P' \leq \frac{D'_{\text{tot}}}{V'} A. \quad (4)$$

To get the inequality (1), we look whether there exists a mapping $v \rightarrow v'$ such that

$$\frac{D_{\text{tot}}}{V} \geq \frac{D'_{\text{tot}}}{V'}. \quad (5)$$

For this purpose let us first investigate the dissipation function D , which is (see [6], p. 50)

$$D = \frac{M_0}{2} (|\kappa_1| + |\kappa_2| + |\kappa_1 + \kappa_2|), \quad (6)$$

where κ_1 and κ_2 are the principal rates of curvature associated with the velocity field v . For domains of (positive) elliptic and parabolic curvature corresponding to the regimes A , AB and AF of the yield hexagon we have

$$D = M_0(\kappa_1 + \kappa_2) = 2M_0H = -M_0 \nabla^2 v, \quad (7)$$

where H is the rate of the mean curvature, and v is counted positive when directed downwards. On the other hand we may write for all regimes

$$D \geq M_0 \max |\kappa_i|, \quad D \geq 2M_0H. \quad (8)$$

We shall later use the fact that (8) is valid at every point of the field v , even if hinges occur. A hinge line may be considered as a narrow strip, where one of the rates of curvature is very large and the other finite.

Finally, we note from (8) and Green's formula, that $v \geq 0$ everywhere, since a domain with $v < 0$ can be removed by $v^* = 0$, thus diminishing $D_{\text{tot}}/\iint v dA$, so that $v < 0$ cannot be the actual field.

Denote now by C_ρ the contour line $v = \rho$ of the surface $v = v(x, y)$ (Fig. 2). We note that C_ρ may consist of several branches. Let κ_t be the rate of curvature tangential to the contour line C_ρ , κ the curvature of the contour line in its plane, $d\varphi = \kappa ds$ the increment of the angle of the tangent at C_ρ , when a point of v moves an increment ds on C_ρ , and finally let $\partial/\partial n$ denote differentiation normal to C_ρ into its "interior", i.e. the direction of increasing ρ . We integrate the dissipation function D over an infinitesimal strip between C_ρ and $C_{\rho+d\rho}$, which gives, by using (8),

$$\begin{aligned} dD_{\text{tot}} &\geq M_0 \oint_{C_\rho} \max |\kappa_i| dn ds \geq M_0 \oint_{C_\rho} |\kappa_t| dn ds \\ &= M_0 \oint_{C_\rho} |\kappa| \left| \frac{\partial v}{\partial n} \right| dn ds \geq M_0 \int_0^{2\pi} d\rho d\varphi = 2\pi M_0 d\rho. \end{aligned} \quad (9)$$

The equality sign in the third inequality of (9) holds, when C_ρ is convex and consists only of one branch.

After these preparations we now specify the mapping of v in two steps as follows: $v(G) \rightarrow v''(G')$, $v''(G') \rightarrow v'(G')$. The first transformation is the so-called *Schwarz' symmetrization* (see [4], p. 190 or [5]). Let C''_ρ be the contour line of v'' , which corresponds to C_ρ (same height ρ), and let A_ρ and A''_ρ be the areas, which C_ρ and C''_ρ "surround" (increasing ρ). Schwarz' symmetrization is then defined as follows

I. C''_ρ is a circle, concentric in G' .

II. The areas A_ρ and A''_ρ are equal.

It is easy to see, that Schwarz' symmetrization does not change the volume V . However D_{tot} might be, at least in certain cases, increased. We introduce therefore a further transformation. The velocity field v'' consists, because of the rotational symmetry, of several ring-shaped circular zones of elliptic, parabolic and hyperbolic rate of curvature (Fig. 3). We replace the body between v'' and the plane of G' by its *convex hull* (surface v' indicated by the dotted lines in Fig. 3), which has only elliptic and parabolic curvature.

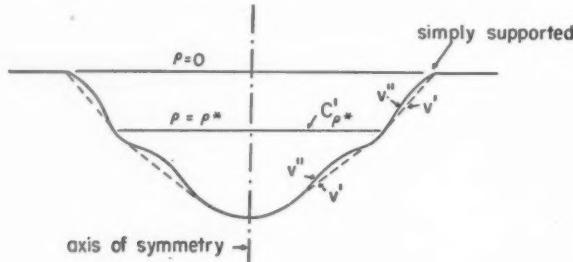


FIG. 3.

Let C'_{ρ^*} be the smallest contour circle, such that outside C'_{ρ^*} no elliptic curvature exists. The transformation has then the following properties

a) The volume is increased: $V' \geq V''$.

b) C'_{ρ^*} is on v' and on v'' ;

$$A'_{\rho^*} = A''_{\rho^*} = A_{\rho^*}.$$

c) $\frac{\partial v'}{\partial n'} = \frac{\partial v''}{\partial n''}$, for $\rho = \rho^*$;

$$\frac{dA'_\rho}{d\rho} = \frac{dA''_\rho}{d\rho} = \frac{dA_\rho}{d\rho}, \text{ for } \rho = \rho^*.$$

As outside C'_{ρ^*} , $D' = M_0 \kappa'_t$, we shall have the equality signs in (9) there. As on the other hand the inequalities (9) are valid everywhere in G , we obtain

$$D_{\text{tot}} \geq D'_{\text{tot}}, \text{ for } \rho < \rho^*. \quad (10)$$

"Inside" C_{ρ^*} , the second inequality (8) can be applied, which gives

$$D_{\text{tot}/\rho \geq \rho^*} \geq -M_0 \iint_{\rho \geq \rho^*} \nabla^2 v \, dA = M_0 \oint_{C_{\rho^*}} \frac{\partial v}{\partial n} \, ds, \quad (11)$$

and inside C'_{ρ^*} we may apply Eq. (7) (regimes A, AB)

$$D'_{\text{tot}/\rho \geq \rho^*} = -M_0 \iint_{\rho \geq \rho^*} \nabla^2 v' dA' = M_0 \oint_{C'_{\rho^*}} \frac{\partial v'}{\partial n'} ds'. \quad (12)$$

As $\partial v / \partial n \geq 0$, $\partial v' / \partial n' \geq 0$, per definition of the contour lines, there exists an important inequality between the last terms in (11) and (12), which is actually the key point of the proof. It follows namely from *Schwarz' inequality* for $\partial v / \partial n$ and $(\partial v / \partial n)^{-1}$ and from the *isoperimetric inequality* $4\pi A_{\rho^*} \leq l_{\rho^*}^2$, where l_{ρ^*} is the length of C_{ρ^*} (see [4], p. 234)

$$\oint_{C_{\rho^*}} \frac{\partial v}{\partial n} ds \geq \frac{4\pi A_{\rho^*}}{\left| \frac{dA_{\rho}}{d\rho} \right|_{\rho=\rho^*}} = \oint_{C'_{\rho^*}} \frac{\partial v'}{\partial n'} ds'. \quad (13)$$

Therefore we obtain

$$D_{\text{tot}} \geq D'_{\text{tot}}, \quad \text{for } \rho \geq \rho^*. \quad (14)$$

From (14) and (10) we conclude, that D_{tot}/V , and also P , are not increased by the mapping $G \rightarrow G'$, and as $6\pi M_0$ is actually the limit load of the circular plate (see [6], p. 55), inequality (1) is proved.

It remains to show that the equality sign in (1) is valid only in the case of the circular plate. The equality sign in (13) holds only when C_{ρ^*} is a circle, and when $\partial v / \partial n$ is constant. The equality signs in the two last inequalities of (9) hold, when every contour line outside C_{ρ^*} consists of one convex branch and is a line of principal curvature, which means that $\partial^2 v / \partial n \partial s = 0$. As one of them, namely C_{ρ^*} , is circular, they must all be circular; therefore, the edge is a circle.

3. v. Mises' yield condition. If one takes v. Mises' yield criterion instead of Tresca's condition, the limit load is not diminished, because the ellipse surrounds the hexagon [7] (Fig. 1). Therefore the inequality (1) remains true

$$P \text{ (v. Mises)} \geq 6\pi M_0. \quad (15)$$

However the inequality is not isoperimetric.

4. Minimum weight design of a sandwich plate. As there is a certain duality between analysis and design problems [8], we expect also an isoperimetric inequality in the latter case. However the result is less useful, because a bound for the minimum volume, for example, does not help in finding the actual design. Nevertheless let us look at a sandwich plate of variable thickness h of the sheets, but constant thickness H_0 of the core, with a homogeneous material obeying Tresca's yield criterion. The yield moment is given by

$$M_0 = \sigma_0 h H_0, \quad (16)$$

where σ_0 is the yield stress. Looking for a statically admissible stress field with regime A of the hexagon (Fig. 1), Prager [9] has shown that h must satisfy the equation

$$\nabla^2 h = -\frac{p}{\sigma_0 H_0}, \quad (17)$$

and he mentioned the analogy between this problem and that of the membrane. The

volume of the sheets

$$V_{\text{statically admissible}} = 2 \iint_A h \, dA \quad (18)$$

corresponds to the torsional rigidity, if we take the torsion analogy instead of the membrane analogy. For the torsional rigidity the isoperimetric inequality has been proved long ago; thus we may write, by using the first theorem of limit analysis,

$$V \leq V_{\text{statically admissible}} \leq V_{\text{circle}} = \frac{\rho A^2}{4\pi\sigma_0 H_0}. \quad (19)$$

5. Steiner's symmetrization. In the case of a very long but narrow domain G , the isoperimetric inequality gives a very bad bound. Steiner's symmetrization (Fig. 4, see also [10]) does not change G so much as Schwarz' symmetrization, if the axis is chosen conveniently. Steiner's symmetrization, which increases the torsional rigidity, therefore

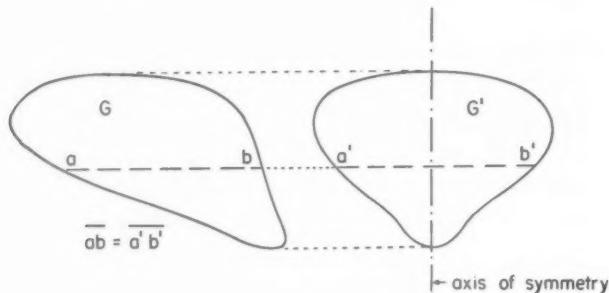


FIG. 4. Steiner's symmetrization.

increases also the volume of a design given in (18), and probably diminishes (or leaves constant) the limit load of a simply supported, uniformly loaded plate. Unfortunately, we are not able to prove the last of these two statements, which would be useful.

BIBLIOGRAPHY

1. H. G. Hopkins and W. Prager, The load carrying capacities of circular plates, *J. Mech. Phys. Solids* **2**, 1-13 (1953)
2. R. M. Haythornthwaite and R. T. Shield, A note on the deformable region in a rigid plastic structure, *Brown University Tech. Rep. C11-26* (1957)
3. D. C. Drucker, W. Prager and H. J. Greenberg, Extended limit design theorems of continuous media, *Quart. Appl. Math.* **9**, 381-389 (1952)
4. G. Polya and G. Szegö, *Isoperimetric inequalities in mathematical physics*, Princeton (1951)
5. H. A. Schwarz, *Gesammelte Math. Abhandlungen*, vol. 2, Göttingen, 327-335, 1890
6. W. Prager, *Probleme der Plastizitätstheorie*, Basel, 1955
7. R. Hill, A note on estimating yield-point loads in a plastic-rigid body, *Phil. Mag.* **43**, 353-355 (1952)
8. D. C. Drucker and R. T. Shield, Design for minimum weight, *Proc. 9th Intern. Congr. Appl. Mech.*, Brussel, 1956
9. W. Prager, Minimum weight design of plates, *De Ingenieur* **48** (1955)
10. J. Steiner, *Einfache Beweise der isoperimetrischen Hauptsätze*, Werke II, Berlin, 75-91, 1882

DETERMINATION OF UPPER AND LOWER BOUNDS FOR SOLUTIONS TO LINEAR DIFFERENTIAL EQUATIONS*

BY R. M. DURSTINE AND D. H. SHAFFER (*Westinghouse Research Laboratories, Pittsburgh, Penn.*)

Analytic upper and lower bounds may be constructed on the solutions to linear differential systems of a certain kind, as described in this paper. The principal requirement of the method presented here is that two functions satisfying the boundary conditions placed on the system—and certain other conditions—must first be constructed. The bounds are then generated as linear combinations of these two functions. The development proceeds as follows.

Consider a linear differential equation

$$L(u) + \phi = 0$$

defined in a suitable region D with boundary B , and subject to the linear conditions

$$M_j(u) = \lambda_j \quad \text{on } B, \quad j = 1, 2, \dots, q,$$

where L and M_j are homogeneous linear differential operators, and q is appropriate to the particular problem. Here ϕ is taken as a known piecewise continuous function in D . The functions λ_j are known on B .

Now select a function w which has the following properties:

- 1) w is piecewise continuous in the highest derivatives appearing in L ;
- 2) $M_j(w) = \lambda_j$ on B , $j = 1, 2, \dots, q$.

Then define ϵ as

$$L(w) + \phi = \epsilon.$$

If $v = u - w$, then

$$L(v) = -\epsilon \quad \text{in } D$$

$$M_j(v) = 0 \quad \text{on } B; \quad j = 1, 2, \dots, q.$$

It is assumed that the original system was such that the above problem has a unique solution, and furthermore possesses a Green's function which does not change sign at any point of D . We have then

$$v = \int_D G(x, \xi) \epsilon(\xi) d\xi,$$

where x and ξ represent points of the space in which we are working, regardless of dimension.

Let w_1 and w_2 be two functions, each satisfying the conditions previously imposed on w . Corresponding to these are two functions ϵ_1 and ϵ_2 . It is required in addition that

- 1) ϵ_2/ϵ_1 be continuous in D ;
- 2) ϵ_1 not change sign in D ;
- 3) one of the following inequalities holds

$$1 > M \geq \epsilon_2/\epsilon_1 \geq m$$

or

$$M \geq \epsilon_2/\epsilon_1 \geq m > 1.$$

*Received June 26, 1957.

We have then

$$\frac{u - w_2}{u - w_1} = \frac{\int_D G(x, \xi) \epsilon_2(\xi) d\xi}{\int_D G(x, \xi) \epsilon_1(\xi) d\xi}.$$

Application of a mean value theorem to the right hand side of this expression gives

$$\frac{\int_D G(x, \xi) \epsilon_2(\xi) d\xi}{\int_D G(x, \xi) \epsilon_1(\xi) d\xi} = \frac{\epsilon_2[\xi_1(x)]}{\epsilon_1[\xi_1(x)]},$$

where ξ_1 is an interior point of D , and clearly a function of x . For brevity we denote $\epsilon[\xi_1(x)]$ by $\delta(x)$. Thus we have

$$\frac{u - w_2}{u - w_1} = \frac{\delta_2}{\delta_1}.$$

Rearrangement of this formula gives

$$u = w_1 + (w_1 - w_2) / \left(\frac{\delta_2}{\delta_1} - 1 \right).$$

If the exact forms of δ_1 and δ_2 were known, this would give the solution to the original problem. However, even though these forms are not known, the ratio is, by hypothesis, bounded.

Since ϵ_2/ϵ_1 is bounded away from unity in D it follows that $w_1 - w_2$ is always of one sign in D , as will be determined by the sign of G . So the second term in the above expression for u may be bounded in D by replacing the denominator by its minimum and maximum values. Define u_1 and u_2 by

$$u_1 = w_1 + (w_1 - w_2)/(M - 1)$$

$$u_2 = w_1 + (w_1 - w_2)/(m - 1).$$

The true value of u must lie between u_1 and u_2 .

The "efficiency" of these bounds may be defined as the spread between them, given by

$$S = |w_1 - w_2| \left[\frac{M - m}{(m - 1)(M - 1)} \right].$$

This formula shows the curious fact that very good bounds may be obtained from functions w_1 and w_2 which are themselves poor approximations to u .

As one example of this method consider the Bessel equation

$$u'' + u'/x + 4u = 0$$

in $0 < x < 1$, subject to $u(0) = 1$. We use the $0 < x < 1$ range to satisfy the conditions imposed on the Green's function. The solution to this problem is $u = J_0(2x)$. As approximating functions we take

$$w_1 = 1$$

$$w_2 = 1 - 1.1167x^2 + .3570x^3.$$

Associated with these are

$$\epsilon_1 = 4$$

$$\epsilon_2 = - .467 + 3.213x - 4.467x^2 + 1.428^3.$$

From these we compute

$$m = - .11675$$

$$M = .05125$$

and hence

$$1.054w_2 - .0540 \leq J_0(2x) \leq .8955w_2 + .1045$$

which produces a rather accurate pair of bounds, particularly near the origin. The function w_2 was of course itself a good approximation to $J_0(2x)$, chosen by another approximation scheme.

As another example consider

$$\Delta\psi = -2$$

in the region interior to $x^2/a^2 + y^2/b^2 = 1$ subject to $\psi = 0$ on the boundary. Choose

$$w_1 = 1 - x^2/a^2 - y^2/b^2$$

$$w_2 = 0.$$

The conditions requisite for the application of this method are fulfilled and we have

$$\epsilon_1 = 2 - 2/a^2 - 2/b^2$$

$$\epsilon_2 = 2.$$

Since both errors are constant, it follows that the precise form of δ_2/δ_1 is known. Hence instead of bounds we obtain the exact solution

$$\psi = (a^2b^2 - b^2x^2 - a^2y^2)/(a^2 + b^2).$$

This is the classical problem of the torsion of an elliptic cylinder.

It should be noted in passing that the results presented here are obtainable under the weaker assumption of only piecewise continuity for ϵ_2/ϵ_1 .

BOOK REVIEWS

(Continued from p. 306)

The hypercircle in mathematical physics. By J. L. Synge. Cambridge University Press, Cambridge, 1957. xii + 424 pp. \$13.50.

In 1947, the author and W. Prager collaborated on a paper entitled, *Approximations in elasticity based on the concept of function space*, in which the "method of the hypercircle" was developed to furnish approximate solutions to boundary value problems of elasticity theory. At the time of this research, the reviewer was privileged to read the correspondence between the authors. It is especially gratifying, therefore, to read the present book which represents a summation of over a decade of work and thought during which the ideas developed and new applications materialized.

The hypercircle method is one of functional approximation in Hilbert space. Given a boundary value problem, a function space is defined with an appropriate metric. The solution of the problem is characterized as the intersection of two orthogonal linear subspaces. This is typified in the Dirichlet problem where one of the subspaces consists of all vector fields obtained as gradients of functions satisfying the boundary conditions, the other subspace consists of all divergence free vector fields. The metric is the Dirichlet integral. By choosing finite sets of particular vectors one constructs a finite dimensional subspace in each of the given subspaces. Minimizing the metric distance between the subspaces determines a pair of vectors, termed "vertices." The solution of the problem lies on a hypercircle having the segment joining the vertices in the space as diameter. This gives upper and lower bounds for the norm of the solution. Knowing the center and radius of this hypercircle, one takes the function associated with the center as an approximation to the unknown solution with error measured in the sense of the metric by the radius. Increasing the dimension of the subspaces reduces the error and improves the approximation.

In certain problems, such as determining the torsional rigidity of a prismatic bar the norm of the solution function is itself of interest. This can be accurately determined with precise error estimates. The function associated with the center of the hypercircle can be evaluated to yield pointwise approximation in the physical domain to the solution, e.g. in the torsion problem to yield approximate stress and warping values. Although the pointwise accuracy of the approximation at any stage is unknown, convergence in the sense of the metric implies pointwise convergence. However, the method can be extended to determine point bounds on the solution and on its derivatives. This is done by bounding the projection of the solution vector on a specially chosen vector which plays the role of a Green's function, although no singularity appears.

Turning to the practical computational problem the author has devised a systematic way of constructing subspaces of increasing dimension so as to contain functions which approximate the solution and its first derivatives to any desired accuracy. These functions are termed "pyramid functions." For a boundary value problem in a domain in two dimensions, such a pyramid function corresponds to a polyhedron constructed over a triangularization of the plane domain with the vertices of the polyhedron located above the vertices in the domain.

The book is divided into three parts. In Part I, elementary properties of linear function space are explored which do not depend on metric.

Part II is devoted to the consequences of the introduction of a positive definite metric and occupies the bulk of the book. In this part the hypercircle method is developed, together with the computational method of the pyramid functions. The results are applied to various boundary value problems of mathematical physics beginning with the Dirichlet and associated problems in potential theory. Particularly detailed calculations are carried out for the problem of determining the torsional rigidity and stress and warping of regular hexagonal and hollow square sections. Excellent bounds are obtained. As examples of mixed boundary value problems the author carries out calculations for the flow of a viscous fluid through a semi-circular channel and the deflection of an elastically supported triangular membrane under uniform pressure. Other boundary value problems for which the machinery of the hypercircle method is set up are those of equilibrium of a three dimensional elastic body and problems associated with the biharmonic equation in hydrodynamics and elasticity.

Part III considers the implications of an indefinite metric in the function space. The hypercircle becomes a "pseudohypercircle" and due to the presence of non-vanishing null vectors in the space, reducing the radius of the hypercircle to zero will not insure convergence of the approximating functions to the solution. However, in those cases where the metric is the difference of two positive-definite met-

trics, as in Minkowski space-time, projections of the hypercircle on the "space-like" and "time-like" subspaces yield certain bounds. Minkowski type metrics arise in problems of vibration. The method of the hypercircle is discussed for the cases of forced elastic and electromagnetic vibrations.

The book is written in the characteristically lucid and interesting style of the author. It is intended for the student as well as the specialist and for the engineer or physicist as well as the mathematician. The pace is leisurely and considerable effort and ingenuity are combined to make the geometry of function space as familiar to the novice as ordinary Euclidian 3-space. There are a good number of interesting problems to be worked out by the reader.

The power of the hypercircle method as seen by the author lies in its wedding of geometry to analysis enabling one to visualize and use intuition to suggest both valid theorems and their proofs. In his introduction, the author refers to other treatments of the problem of bounding solutions of boundary value problems which proceed without diagrams or geometrical ideas. These he recommends to readers who "prefer to take their analysis neat." This reviewer, however, heartily recommends to all readers the stimulating mixture which Professor Syngre has concocted for us.

H. J. GREENBERG

Nonparametric methods in statistics. By D. A. S. Fraser. John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1957. x + 299 pp. \$8.50.

The title of this book is somewhat misleading; the reader who wants a comprehensive collection of nonparametric statistical methods will not find it here. The author's aim seems rather to have been to give a theoretical treatment of those nonparametric techniques which admit some theoretical justification, and in the state of knowledge extant when the book was written this meant excluding mention of many important problems and many useful and commonly employed techniques.

In the first two chapters (almost a half) of the book, the author gives a general introduction to statistical inference which for the most part follows lines of development similar to those employed by E. L. Lehmann in his mimeographed Berkeley notes. Many of the examples here are the standard parametric ones. Thus, the reader will find an introduction to statistical decision theory and such topics as sufficiency, unbiasedness, and invariance in estimation and testing hypotheses. However, the reader who believes the prefatory remark that the calculus and Hoel are the only prerequisites, may find the rapid measure-theoretic excursion of Chapter 1 overwhelming; and the professional mathematician may be a bit disturbed that the measure-theoretic care is dispensed with at the start of the next chapter and that most of the hard and interesting proofs are omitted while those given are usually trivial. The omissions can perhaps be justified on the grounds that these generalities are not the main content of the book.

The last five chapters deal with various nonparametric problems and properties of certain procedures, especially tolerance regions, unbiased estimators, and rank tests. Some of the emphasized topics are treated rather completely, e.g., limit theorems concerning procedures of the last two types just mentioned, as well as the common conditional and run tests. As has been mentioned, many important topics are unfortunately omitted entirely; among these may be mentioned the use of the sample quantiles and sample distribution function in estimation and testing hypotheses. The χ^2 techniques receive only a brief historical mention.

Problems at the end of chapters supplement and extend the text. Within the framework of the latter, they should be instructive to students.

J. KIEFER

An introduction to matrix tensor methods in theoretical and applied mathematics. By Sidney F. Borg. J. W. Edwards, Publisher, Inc., Ann Arbor, Michigan, 1956. 202 pp. \$4.75.

The reviewer could find no excuse for the publication of this book, which contains many examples of erroneous thinking in both mathematics and physics. To give only one example, on page 61 the following statement can be found. "In general, the requirements that a function $u(x,y)$ be continuous at a point (x_0, y_0) are that all partial derivatives be finite and continuous at (x_0, y_0) and that the remainder term of the Taylor Series approach zero as the number of terms increases."

R. T. SHIELD.

Economic models: an exposition. By E. F. Beach. John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1957. xi + 227 pp. \$7.50.

For the non-genius, a knowledge of mathematical techniques is a sine-qua-non for work in economic theory; and it is highly desirable for the economic theorist to learn some pure mathematics. To be sure, a genius can get by on his high school plane geometry; but I venture the guess that the complexity of problems is growing at a faster rate than the mental capacity of human beings. As a result, it will soon be necessary for even a genius to learn mathematical techniques. Professor Beach realizes that "many students become interested in economic theory only after they have left the study of mathematics at an early stage some years previously. It is rather difficult for them to return to the study of mathematics when they are rather fully occupied with the requirements of advanced degrees." When the student finishes reading this book, Professor Beach believes, he should know whether he wants to delve deeper into the literature of mathematical economics and statistics.

An applied mathematician will not learn economics or mathematical economics from this book. If he is innocent of statistics, he could read part two with profit. The author provides a lucid and interesting introduction to sampling theory, simple and multivariate regression theory and the book concludes with a very useful chapter on the Cowles Commission's Simultaneous Equations Approach. Koopmans' famous article on Identification Problems in Economic Model Construction is discussed in this chapter, and it can be read with profit by many students. A bibliography is provided for those whose appetites for statistics have been whetted.

The cardinal deficiency of this book is that it fails to concentrate on a few economic problems and show how mathematical techniques enable the economists to derive a more profound understanding of the economic processes at work. For example, Beach tells the student that Hicks and Modigliani enabled economists to understand Keynes' contribution more fully, when they translated his model into mathematical terms. Beach then writes down their equations, but neglects to show the student how the mathematical model is more revealing than the literary model. Hicks translated Keynes into mathematics and then described his mathematical results in terms of two dimensional geometry; then, Hicks translated the geometry back into literary economics. Thereby, the mathematical reader, the semi-literate mathematical reader and the literary economist could benefit from Hicks' mathematics. Beach, on the other hand, merely writes down Hicks' equations and leaves it there. In the beginning of chapter three Beach interweaves algebra and geometry (familiar to economists) in a lucid and interesting manner. The student is lead back and forth between familiar geometry and unfamiliar algebra; and is thereby taught some algebra. Unfortunately, this method of exposition is not continued.

As the book progresses the exposition deteriorates in quality. The student is given an introduction to the derivative and is taught that $Dx^n = nx^{n-1}$. Then twenty pages later the author begins to solve differential equations, using the term integration and techniques of integration; but he never taught the student anything about an integral or integral calculus. A student who was familiar with integral calculus would not need to study this book; and one unfamiliar with integrals would be mystified by his chapter on continuous dynamic models.

Parts of this book are excellent introductions to the use of mathematical techniques in economics. Moreover, it is lucid and interesting. I have no doubt that Beach's major purpose will be satisfied: the student will learn whether or not he wants to learn mathematical economics.

J. L. STEIN

Mathematical analysis—a modern approach to advanced calculus. By Tom M. Apostol. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1957. xii + 553 pp. \$8.50.

The author properly describes his book by the statement in the preface, ". . . most of the topics which usually fall under the heading 'Advanced Calculus' are treated in this book. The author's aim has been to provide a development of the subject matter which is honest, vigorous, up-to-date, and, at the same time not too pedantic." The book is designed primarily as a text for (pure) mathematics students but is at a level of difficulty somewhat below the usual advanced analysis. The book is written in a clear leisurely style with ample foreshadowing and motivation of the developments even though there is practically no discussion of applications.

G. NEWELL

k,

ie
e,
of
n
y
at
e-
ne
ve

c.
l-
l-
n
e

d
d
-
e
-
s
-
l,
r
g
s

n
e
t
l
y

.

3
1
t
3
1
1

WILEY

BOOKS

*Ready Now—Vol. I**In Press—Vols. II and III*

HANDBOOK OF AUTOMATION, COMPUTATION, AND CONTROL

Prepared by a staff of 106 specialists

Edited by EUGENE M. GRABBE, SIMON RAMO, and DEAN E. WOOLDRIDGE, all of the Ramo-Wooldridge Corp. Written and edited with systems engineering emphasis in mind, the handbook covers material of direct use to all levels of technical personnel, in all areas of control. The major objective is to provide practical data for research, development, and design in the fields of feedback control, computers, data processing, control components, and control systems.

Throughout, the stress is on new techniques and components for designing and building digital devices, making measurements, and developing control systems. Much of this material has not been available previously in handbook form.

Each chapter starts with definitions and descriptions, then builds up to more complex details. Thus the handbook provides both an *elementary frame of reference* for the novice and *advanced data* for the expert.

Volume I—CONTROL FUNDAMENTALS

- Covers aspects of mathematics, *as applied to control*.
- Includes a compilation of numerical analyses for digital computers—the latest techniques and comparisons of different techniques.
- Presents self-contained treatments of feedback control theory and of operations research.
- Gives essential material on information theory and data transmission.

1958.**Over 1000 pages.****\$17.00.**

Volume II—COMPUTERS AND DATA PROCESSING. In Press.

Volume III—SYSTEMS AND COMPONENTS. In Press.

Write for an examination copy of Vol. I,
and reserve examination copies of Vols. II and III.

JOHN WILEY & SONS, Inc. 440 Fourth Ave. New York 16, N. Y.

WILEY

BOOKS



Now ready . . .

SURVEYS in APPLIED MATHEMATICS

These survey articles, work on which was initiated by the Editorial Office of *Applied Mechanics Reviews* at the request of the Office of Naval Research, have the primary objective of reviewing recent mathematical progress in the fields which they cover. The surveys cover the international literature, and pay special attention to Russian contributions which are not yet readily available in the English language literature. The articles are aimed at a broad, mathematically literate audience looking for an up-to-date account of modern progress in applied mathematics and an appraisal of future promising research directions. The authors, without exception, are internationally recognized authorities in the areas of applied mathematics dealt with by their surveys.

The Series . . .

Vol. I—Elasticity and Plasticity

By J. N. GOODIER and P. G. HODGE, Jr. 1958. In press

Vol. II—Dynamics and Nonlinear Mechanics

By E. LEIMANIS and N. MINORSKY. 1958. In press

Vol. III—Mathematical Aspects of Subsonic and Transonic Gas Dynamics

By L. BERS. 1958. In press

Vol. IV—Some Aspects of Analysis and Probability

By I. KAPLANSKY, E. HEWITT, M. HALL, JR., and R. FORTET. 1958. In press

Vol. V—Numerical Analysis and Partial Differential Equations

By G. E. FORSYTHE and P. C. ROSENBOOM. 1958. In press

Send for examination copies.

JOHN WILEY & SONS, Inc. 440 Fourth Ave. New York 16, N. Y.



SUGGESTIONS CONCERNING THE PREPARATION OF MANUSCRIPTS FOR THE QUARTERLY OF APPLIED MATHEMATICS

The editors of "Quarterly of Applied Mathematics" request the following directions for the preparation of manuscripts. These directions have been drawn up with a view toward convenience of manuscript correspondence, avoiding the return of papers for changes, and reducing the charges made for "author's corrections."

Manuscripts: Papers should be submitted in original type, either on one side only of white paper sheets and be double or triple spaced with wide spacings. Marginal instructions to the printer should be written in pencil so distinctly that they clearly from the body of the text.

The papers should be submitted in final form. Only typographical errors may be corrected in proof; composition errors for an error detected from the manuscript will be passed on to the editor.

Titles: The title should be brief but express adequately the subject of the paper. The name and initials of the author should be written at the front of the title page and hence will be omitted. The name of the organization with which the author is connected should be given in a separate line to follow his name.

Mathematical Works: As far as possible, formulae should be typeset in Greek letters and other symbols not available in the typesetter's alphabet carefully inserted in ink. Manuscripts containing pencil material other than proofs of figures must be typed in ink.

Figures: Figures should be submitted in final form. All dimensions and scales should be clearly indicated on the figure. Axes should be labeled with the name of the axis and its unit. The origin of the coordinate system should be marked with zero. If, and when, a dimensionless quantity, the ratio f , and the prime ($'$), between length and a length and ℓ , respectively, are used, and if f and ℓ are not explicitly mentioned, it should be clearly indicated.

Dots, bars, and other markings to be set off by dots should be uniformly spaced because they require easily hand-composition. In their usual markings (such as John's initials) which follow the letter should be used.

Square roots should be written with the exponent 1, not greater than the sign $\sqrt{}$.

Complicated exponents and subscripts should be avoided. Any complicated expression that occurs frequently should be given once by a symbol ρ .

For convenience when setting up equations, the exponent or symbol α should be used, particularly if such exponents appear in the body of the text. Thus,

$$\cos (ax + b)^\alpha \text{ is preferable to } \cos^\alpha (ax + b)$$

Fractions in the body of the text and fractions occurring in the numerators or denominators of fractions should be written with their取消符. Thus,

$$\frac{\cos (ax + b)}{\cos (ax/c)} \text{ is preferred to } \frac{\cos \frac{ax}{c}}{\cos \frac{ax+b}{c}}$$

In many instances the use of negative exponents permits saving of space. Thus,

$$\int u^2 dx \text{ is preferable to } \int \frac{du^2}{u} dx$$

When the bracketed fraction is composed of two or three terms, it is better to write the right parenthesis in position of the first term and the left parenthesis in position of the last term. To indicate the extent of the bracketed fraction, the parentheses should be placed as follows:

$$(a + bx) \cos t \text{ is preferable to } \cos t (a + bx).$$

In the derivation of formulas, the use of brackets is to be avoided. Instead, the use of parentheses is to be preferred. In particular, when dealing with polynomials of partially known coefficients, such as $(1 + x)^n$, the brackets should be avoided, and the parentheses should be used.

For example, the formula $(1 + x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \dots$ is to be preferred to $(1 + x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \dots$

Cuts: Carelessly placed cuts, especially when they occur in the middle of a line, are to be avoided. It is better to break the line at the end of the cut and to start the next line at the beginning of the cut. This is particularly important in the preparation of tables. The use of the carriage return at the end of the line is to be avoided, as it causes unnecessary blank space.

The following examples show the way in which cuts are to be avoided. The first example, from the Quarterly, vol. 2, is to be preferred to the second, in which the carriage return is used at the end of the line.

Author's Note: It is suggested that the author supply the printer with a copy of the typescript, so that he may be able to correct any errors in the typesetting.

The use of the carriage return at the end of the line is to be avoided, as it causes unnecessary blank space.

Footnotes: If the author wishes to add a note to the text, it should be done in the margin, and the note should be numbered.

Afterword: If the author wishes to add a note to the notes, it should be done in the margin, and the note should be numbered.

Footnote: If the author wishes to add a note to the notes, it should be done in the margin, and the note should be numbered.

Footnote: If the author wishes to add a note to the notes, it should be done in the margin, and the note should be numbered.

